Statistical Models in Prognostic Modelling With Many Skewed Variables and Missing Data: A Case Study in Breast Cancer

Thesis submitted for the degree of Doctor in Philosophy

Mohammad Reza Baneshi

College of Medicine and Veterinary Medicine

Edinburgh University

May 2009

Abstract

Prognostic models have clinical appeal to aid therapeutic decision making. In the UK, the Nottingham Prognostic Index (NPI) has been used, for over two decades, to inform patient management. However, it has been commented that NPI is not capable of identifying a subgroup of patients with a prognosis so good that adjuvant therapy with potential harmful side effects can be withheld safely.

Tissue Microarray Analysis (TMA) now makes possible measurement of biological tissue microarray features of frozen biopsies from breast cancer tumours. These give an insight to the biology of tumour and hence could have the potential to enhance prognostic modelling. I therefore wished to investigate whether biomarkers can add value to clinical predictors to provide improved prognostic stratification in terms of Recurrence Free Survival (RFS).

However, there are very many biomarkers that could be measured, they usually exhibit skewed distribution and missing values are common. The statistical issues raised are thus number of variables being tested, form of the association, imputation of missing data, and assessment of the stability and internal validity of the model.

Therefore the specific aim of this study was to develop and to demonstrate performance of statistical modelling techniques that will be useful in circumstances where there is a surfeit of explanatory variables and missing data; in particular to achieve useful and parsimonious models while guarding against instability and overfitting. I also sought to identify a subgroup of patients with a prognosis so good

that a decision can be made to avoid adjuvant therapy. I aimed to provide statistically robust answers to a set of clinical question and develop strategies to be used in such data sets that would be useful and acceptable to clinicians.

A unique data set of 401 Estrogen Receptor positive (ER+) tamoxifen treated breast cancer patients with measurement for a large panel of biomarkers (72 in total) was available. Taking a statistical approach, I applied a multi-faceted screening process to select a limited set of potentially informative variables and to detect the appropriate form of the association, followed by multiple imputations of missing data and bootstrapping. In comparison with the NPI, the final joint model derived assigned patients into more appropriate risk groups (14% of recurred and 4% of non-recurred cases). The actuarial 7-year RFS rate for patients in the lowest risk quartile was 95% (95% C.I.: 89%, 100%).

To evaluate an alternative approach, biological knowledge was incorporated into the process of model development. Model building began with the use of biological expertise to divide the variables into substantive biomarker sets on the basis of presumed role in the pathway to cancer progression. For each biomarker family, an informative and parsimonious index was generated by combining family variables, to be offered to the final model as intermediate predictor. In comparison with NPI, patients into more appropriate risk groups (21% of recurred and 11% of non-recurred patients). This model identified a low-risk group with 7-year RFS rate at 98% (95% C.I.: 96%, 100%).

I then elaborated these methods with investigation of elements of procedure: screening and imputation of missing data, and appropriate form of association.
'Median substitution' method provided results comparable to sophisticated multiple imputation technique, probably due to low rate of missing data. Furthermore, submission of all of the biomarkers to the model slightly changed composition of the final model in terms of selection of variables and resulted in inflated S.E.'s.

Regarding the form of association, the superiority of data-driven techniques over pre-specified methods was confirmed.

By performing a multifaceted screening followed by multiple imputations and bootstrapping (to check both the stability of form of association and reliability of inclusion across models), I developed a methodology which has the potential for future application in all medical areas. Furthermore, powerful predictive biomarker tools have been proposed which promise to increase understanding and prevention of breast cancer progression and which provide a significant potential improvement over conventional NPI risk stratification. However, models developed warrant validation in a larger cohort, ideally with longer follow-up.

Declaration form

The data set analysed in this thesis was collected under the direction of Professor John Bartlett and his colleagues, firstly at Glasgow University when Professor Bartlett worked there, and latterly at the University of Edinburgh,

The statistical ideas for the PhD study were mine. However my regular meetings with my supervisors and the clinical collaborators of this study, to address the statistical approaches useful in the field of time-to-event clinical outcome data, and in presentation of results in a way meaningful to a clinical audience, will have contributed to the evolution of some ideas. For the 'Biologically informed' modelling process, Professor Bartlett specified the division of biomarker variables into substantive sets (based on his unique specialised expertise in the biology of cancer progression and knowledge of the research field). In addition, his contributions to drafting the biological discussion of a clinically-orientated manuscript (derived from my PhD research, which is ready for publication), have been fed back into the thesis.

Although the composition of the thesis has been entirely my own work, my supervisors, Dr Pamela Warner and Dr Niall Anderson, have provided extensive help to me in developing my communication in English of statistical research and reflection, and in many instances have provided rephrasing of sentences or paragraphs. Nevertheless, they support my declaration that this thesis is my personal effort. The work has not been submitted for any other degree or qualification.

Signed:

Table of content

List of Tables

List of figures

Dedication

I dedicate this thesis to my wife, Farzaneh and my son, Kiamehr. Without their caring support and understanding the completion of this work would not have been possible.

Acknowledgments

I should dedicate my grateful acknowledgements to my supervisors, Dr Pamela Warner and Dr Niall Anderson, for their support, patience, and professional guides. My supervisors provided detailed guidance, insightful comments and suggestions, and encouragement throughout in all stages of this project.

Also I should thank Professor Robin Prescott who facilitated my access to a unique breast cancer data set and Professor John Bartlett who provided the data to me.

I would like also to convey my thanks to Professor John Bartlett, Dr Jill Kerr, Dr Wilma Jack, and Professor David Cameron for their enthusiasm and feedback during analysis of the data set.

This research is supported by Iranian Ministry of Health and Medical Education, and also Kerman Medical University.

Abbreviations

AIC	Akaike's information criterion (AIC)
AKT	AKT biomarker set
BAD	BAD biomarker set
B.E.	Backward Elimination variable selection
BGVS	Biologically Guided Variable Selection
CART	Classification And Regression Trees
C-C	Complete-Case analysis
CI	Confidence Interval
C-index	Harrell's concordance index
DCIS	Ductal Carcinoma In Situ
EM	Expectation Maximum
EPV	Event Per Variable
ER+	Estrogen Receptor positive
FP	Fractional Polynomial
FP1 and FP2	First and second degree Fractional Polynomial
GAM	General Additive Models
HER	HER biomarker set
HR	Hazard Ratio
IDC	Invasive Ductal Carcinoma
K-M	Kaplan-Meier survival curves
MAPK	MAPK biomarker set
MAR	Missing At Random

Max	Maximum
MCAR	Missing Completely At Random
MFP	Multivariable Fractional Polynomial
MICE	Multivariate imputation via chain equations
Min	Minimum
MNAR	Missing Not At Random
MTOR	MTOR biomarker set
NPI	Nottingham Prognostic Index
NPI^{q4}	Division of patients into 4 risk groups applying cut offs at quartiles of NPI
NPI^{std3}	Division of patients into 3 risk groups applying standard cut offs at NPI
NPI^{std4}	Division of patients into 4 risk groups applying published cut offs at NPI
NRI	Net Reclassification Index
OS	Overall Survival
PCA	Principal Component Analysis
PgR	PgR family set
PSEP	Prognostic SEParation
Q1, Q2, Q3	Bottom, middle, and top quartiles
RAS	RAS family set
RCS	Restricted Cubic Spline
recNPI	Recalculated Nottingham Prognostic Index
$recNPI^{q4}$	Division of patients into 4 risk groups applying cut offs at quartiles of recalculated NPI
$recNPI^{std4}$	Division of patients into 4 risk groups using recalculated NPI

RFoT	Recurrence Free on Tamoxifen treatment
RFS	Recurrence Free Survival
S.E.	Standard Error
TMA	Tissue Microarray Analysis
TSM	Tree-based Survival Models
UIVS	Univariately Informative Variable Selection

Chapter 1 GENERAL INTRODUCTION

1.1 Introduction

Clinical trials typically involve collection of patient data at entry and in so far as are possible these data will include variables of potential relevance to the likely cause of the disease under study. These data sets have been a valuable resource in identifying important risk factors for disease course and hence also for risk stratification of patients.

Prognostic models combine key patient characteristics (risk factors) to predict clinical outcomes such as recurrence of cancer. These models are excellent tools to investigate the contribution of variables to disease course, to inform patients about

their likely outcome, to design future studies, and to select treatment paths [Altman DG and Lyman GH, 1998].

A prognostic model, to be useful in practice, should be able to stratify in terms of risk: that is to identify a subset of good and poor prognosis cases that do not and do require further treatments such as chemotherapy [Lee AH and Ellis IO, 2008].

In the case of breast cancer, multiple conventional prognostic candidate variables have been investigated by applying the Cox regression model to derive an index of risk of disease recurrence [Haybittle JL et al., 1982]. In particular, the Nottingham Prognostic Index (NPI) was devised to estimate the risk of recurrence and to classify patients into risk groups [Haybittle JL et al., 1982]. This model uses information on nodal status, tumour grade, and tumour size, and has been widely validated [Todd JH et al., 1987; Galea MH et al., 1992; Balslev I et al., 1994].

The NPI is now central to the risk stratification of patients across the UK. However, it is not capable of identifying a subgroup of patients with a prognosis so good that a decision can be made to avoid treatments with potential harmful side effects [Balslev I et al., 1994]. It has been emphasized that further prognostic factors, besides those used in NPI, are needed to enhance risk stratification [Balslev I et al., 1994].

Referring to the patients who do not require treatment, it has been postulated that ‘it is an inability to identify such patients prior to treatment, rather than an expectation that all patients derive benefit, which drives the treatment of significant number of

breast cancer patients with often aggressive chemotherapy' [Faratian D and Bartlett JM, 2008]. The identification of novel prognostic markers and integration of them in risk prediction is key to the solution of this dilemma [Faratian D and Bartlett JM, 2008].

In recent years data on new potentially informative prognostic variables have become available, in particular using Tissue Microarray Analysis (TMA). The development of automated laboratory techniques has enabled retrospective testing of stored tissue samples in existing cohorts, providing a multitude of potential biological markers for prognostic modelling in breast cancer. Tissue microarray data usually exhibit skewed distributions and their analysis is hampered by missing data.

These TMA variables reflect the biology of a tumour and hence could have the potential to enhance prognostic modelling. In this thesis I will explore the feasibility and methodological aspects of combining biomarker data and clinical variables to develop an enhanced prognostic model providing improved prognostic stratification in terms of Recurrence Free Survival (RFS).

1.2 Overview of the research

The specific aim of this research is to develop and to demonstrate the performance of statistical methods that will be useful in circumstances where there is a surfeit of potential biological tissue microarray data. This is to ascertain which tissue variables are important for outcome so as to enhance understanding of the biology of breast cancer progression, and to achieve useful and parsimonious models, guarding against instability and overfitting.

Other objectives related to the nature of these skewed biomarkers are to find the appropriate form of risk function and to minimise the loss of information due to missing data.

Methods to tackle these challenges are explored and discussed in the process of analysing biological tissue microarray data in breast cancer. The original contribution of this thesis will be to advance understanding of the value of appropriate statistical methods in such circumstances. In addition, it is expected that the findings of the research will reveal new insights into breast cancer aetiology and treatment, and hence lead to better management of patients.

1.3 Structure of this thesis

The remainder of this thesis is structured as follows.

Chapter 2 provides a background to breast cancer. In addition, details of the development of NPI, and its main applications in the literature, are given.

Chapter 3 presents a literature review of the statistical techniques relevant to this project.

Chapter 4 sets out the aims and objectives of this study, and design. An overview of statistical methods relevant to the whole thesis is presented. Specific methods applied are given in relevant results chapters.

Results are then presented in six chapters (Chapters 5 to 10): preliminary work in Chapters 5 and 6, main modelling in Chapters 7 and 8, and enhancement of understanding of methods applied in Chapters 9 and 10.

Chapter 5 describes summary statistics for the variables.

Chapter 6 is devoted to application and also recalculation of NPI. To recalculate the index, new regression coefficients will be calculated with respect to the data set I analysed. Furthermore, effect of categorisation of patients into 3 and 4 risk groups on estimation of event free rates is checked.

In Chapter 7, I will apply alternative screening methods to compare them in terms of detection of form of risk function and selection of univariate potentially informative biomarkers.

In Chapter 8, I will report the process of application of methods to develop multifactorial models with many skewed variables in presence of missing data.

In Chapter 9, models developed in Chapter 8 are challenged by applying alternative approaches to handle missing data.

In Chapter 10, the importance of selection of form of risk function on the composition and performance of the model is addressed.

Chapter 11 provides general discussion of this thesis and suggests priorities for future studies. A set of recommendations for future studies is provided.

Chapter 2 BACKGROUND TO BREAST CANCER, ITS TREATMENT AND RISK PREDICTION

2.1 Introduction

In human body, genes control growth of cells. Healthy cells might become cancerous when abnormal changes happen in genes [Breastcancer.org, 2008f]. In general, 90% of cancers are caused by a genetic abnormality that happen due to aging [Breastcancer.org, 2008c]. Based on 2006 statistics, only 0.5% of cancers registered were in those aged less than 15 years while in 74% age was more than 60 years [Office for National Statistics, 2009].

Over time cancerous cells might invade healthy breast tissues and become a tumour. Cancerous cells can also spread to nearby tissues and enter the bloodstream affecting other parts of the body. Metastatic breast cancer means that the cancerous cells spread to other parts of the body. The extent that initial cancerous cells are spread in the body shows disease stage [Breastcancer.org, 2008f].

Cancer is one of the most major health problems worldwide. In 2002, a quarter of the 11 million new cases of cancer reported worldwide occurred in Europe. In the UK, per year, more than a quarter of a million new cases are diagnosed. Among them, the most prevalent carcinomas (incidence rate) were breast (16%), lung (13%), bowel or colorectal (13%) and prostate (12%) [Cancer Research UK, 2007].

Breast carcinoma, with one million newly diagnosed cases annually, is the most prevalent malignancy among women worldwide, comprising 18% of all female cancers [McPherson K et al., 2000]. The highest rate of breast cancer occur in Northern Europe and North America (101.1 for US and 88.7 for Denmark per 100,000 thousand population) and the lowest rates are in parts of Africa and Asia (19 in Zimbabwe and 18.7 in China) [Cancer Research UK, 2008]. This indicates geographical variation in incidence of this disease.

However, even among European countries, the UK has a high incidence rate (87.2 per 100,000): each year more than 44,000 women are diagnosed with breast cancer [Cancer Research UK, 2008]. As the incidence of breast cancer is high, and five-year survival rates are over 75%, many women are alive who have been diagnosed with

breast cancer. Around 172,000 women are alive in the UK having had a diagnosis of breast cancer [Micheli A et al., 2002].

In the UK, 12400 deaths occurs due to this malignancy per year [Cancer Research UK, 2006]. In Scotland the standardised mortality rate is 62.6 (per 100,000 women). The rate in England and Wales was lower (56.1). The rate in Japan was very low while in Canada very high (21.9 versus 71.1) [McPherson K et al., 2000].

In section 2.2 construction and application of Nottingham Prognostic Index (NPI) for risk stratification in the literature is reviewed. In section 2.3, treatment options and their side effects are given. In section 2.4, the importance of integration of biology to optimise risk prediction and treatment selection is highlighted. An overview is given in 2.5.

2.2 Nottingham Prognostic Index (NPI)

2.2.1 Development of NPI

Currently Nottingham Prognostic Index (NPI) is the gold standard in prognostic method for cases diagnosed with breast cancer. The NPI is considered in conjunction with additional factors such as age and hormonal receptor status to inform management of breast cancer patients across the UK.

The NPI was developed with follow-up data collected from 500 cases with invasive carcinoma treated by simple mastectomy and triple-node biopsy at the Nottingham City Hospital [Haybittle JL et al., 1982]. At the time of analysis, the range of follow-up times was between 1 and 6 years. Of 500 cases recruited, 113 cases were excluded for one or more of the following reasons: 69 cases with lack of data on Estrogen Receptor (ER); 10 cases with no cell reaction score; 11 cases with non-invasive cancer; and 23 cases for a variety of reasons such as operation not being mastectomy or no follow-up possible. The remaining 387 cases were analysed to create the index.

Not all patients were treatment free. After the first 250 cases had been recruited, the publication of Blamey *et al.* caused a change in clinical practice, such that cases judged by Blamey's research to have poor prognosis (those with tumour in either of apical or internal mammary nodes (stage 3), tumour size> 2cm, and grade 2 or 3) should receive chemotherapy [Blamey RW et al., 1979]. This policy was applied to 120 of the patients recruited to the NPI study, the 251st to 370th. However, this policy was then discontinued for the remaining patients up to 500th.

To develop the index, 9 variables were submitted to the multifactorial Cox regression and a full model was fitted: age, menopausal status, tumour size, lymph-node involvement, tumour grade, cellular reaction, presence of sinus histiocytosis in lymph nodes, estrogen receptor (ER), and a binary variable indicating whether adjuvant chemotherapy therapy had been given. This last variable was included because fifteen cases had received chemotherapy.

The only variables making significant contributions to the multifactorial prognostic model were tumour size, lymph-node stage, and grade. Histological grade and stage had 3 levels. In the multifactorial analysis, lymph-node stage showed the strongest association, with a Z-value of 5.29, followed by grade, and tumour size (Z-values 4.56 and 2.92 respectively).

Although nodal status, grade, and tumour size were the only significant variables, no additional regression model was fitted with only these 3 variables. Nevertheless, the risk score was calculated for each patient using only the 3 significant variables in the formula where the multiplier for each variable is the corresponding regression coefficients for the model:

0.17 x Size (cm) + 0.76 x Node {1, 2, 3} + 0.82 x Grade {1, 2, 3}

The authors then applied the risk score in subsets of 298 cases with data on all prognostic factors. Subjects were selected from patients 1-250 and 371-500 thus excluding cases recruited during the period in which poor prognostic cases[1] received chemotherapy.

On the basis that lymph-node stage was the strongest predictor, the index derived was compared with this predictor alone. In the subset studied, the number of patients classified by node stages as 1, 2, and 3 were 154, 95, and 49 respectively. Kaplan-Meier (K-M) survival curves for these node groups were plotted. Cut points were then applied to the index scores derived so as to define three risk groups containing

[1] Poor prognostic cases were defined as having Stage C, size > 2cm, and Grade 2 or 3

the same number of patients (154, 95, and 49), which in their sample was equivalent to 3.65 and 4.5. Comparing K-M survival curves, for these groups compared to lymph node alone, it was seen that the survival of the lowest risk group detected by the index and node stage alone were fairly similar in event free survival. However, the index gave a marginally better discrimination between low and high risk patients.

A second comparison was made between the index and criteria used to define high risk patients, based on Blamey's study [Blamey RW et al., 1979]. Only 25 cases formed the high risk group. The survival curve of these patients was compared with that of 65 patients with the highest index value (≥4.4). Curves were almost identical. However, the index identified a larger number of cases as poor prognosis (65 for index versus 25 based on Blamey's study). Finally, the K-M survival curve of 64 cases with index values below 2.8, were compared with that of the expected survival in a normal population of the same age distribution. The normal population had slightly better survival but very similar. The authors concluded that the risk groups derived were able to define very good low and high risk patients [Haybittle JL et al., 1982].

To improve the ease of application of index, regression coefficients for tumour size, nodal status, and grade were rounded up from 0.17, 0.76, and 0.82 to 0.2, 1, and 1 respectively. The NPI therefore becomes:

NPI = 0.2 x Size (cm) + Nodal status {1, 2, 3} + Grade {1, 2, 3}

In order to produce the same risk groups and similar survival curves, the cut points then had to be shifted in line with the rounding, from 2.8 and 4.4 to 3.4 and 5.4 respectively.

Some years later, the original NPI which was developed on short-term follow-up data, was calculated for all of the 387 cases which had a longer follow-up period and for a new prospective sample containing 320 cases followed-up for 1.7 to 6.5 years (707 cases in total) [Todd JH et al., 1987]. K-M survival curves for original sample (387 cases) and prospective sample (320 cases) were plotted and compared. Survival curves in the corresponding risk groups were fairly similar. Also, combining the two samples together, actuarial 5-years survival of low, intermediate and high risk groups were 88%, 69%, and 22% respectively indicating the ability of NPI to stratify patients into divergent risk groups.

2.2.2 Literature review of applications of NPI

NPI was then validated in several independent samples, as summarised in Tables 2.1 and 2.2. Using Pubmed database, the keyword 'Nottingham' in the title of the paper was searched. This resulted in 470 papers, the majority of them were not relevant to NPI and breast cancer.

Only papers which reported the application of the NPI to stratify patients into risk groups were considered (18 papers in total). Estimated event-free rates in risk groups are summarised in Tables 2.1 and 2.2. Comparison of results is not straightforward as different patient subtypes receiving a variety of treatment regimes were analysed. My

main purpose was to investigate the ability of NPI in identification of low risk patients. Therefore, I reported estimated event-free rate in the lowest risk groups (those with NPI lower than 3.4). For the sake of comparison, estimates in intermediate, and high risk patients are also presented.

Not all papers provided detailed information about number of patients and recurrences in risk groups, follow-up time, and estimated event-free rates. In the case of no report of event-free rates, if survival curves were presented, then figures were judged by that. Furthermore, none of the studies provided information on confidence interval of reported survival rates.

The main findings were as follows. My especial focus is on estimated long-term event free rates (10 years or more) in the low risk groups. However, 6 studies reported short-term rates at 5 years (Table 2.1). Actuarial 5-year survival rate derived from original NPI was 88%. This rate varied from 82% [Sauerbrei W et al., 1997; Coradini D et al., 2001] to 96% [Okugawa H et al., 2005]. Furthermore, 3 studies were performed on node negative breast cancer patients and therefore no patient was assigned into high risk group, so that there were only 2 risk groups [Sauerbrei W et al., 1997; Coradini D et al., 2001; Ring BZ et al., 2006].

Focusing on long-term survival rates (Table 2.2), in the largest studies, nearly 25000 and 10000 patients were recruited [Lundin J et al., 2006; Balslev I et al., 1994]. In both studies, estimated 10-year survival rate was about 80%. Some other studies reported a similar rate [Galea MH et al., 1992; Lundin J et al., 2006]. On the other

hand, in the smallest studies, (n= 82 and 97), a marginally higher survival rate at 10 years (83%) was reported [Sidoni A et al., 2004; Eden P et al., 2004].

Applying the NPI on patients with small primary breast cancer the highest 10-year event-free rate was reported at 88% [Kollias J et al., 1999] which was also reported in the longest follow-up study [D'Eredita' G et al., 2001]. Estimated event-free rate for high-risk cases in the latter study was also higher than most of the other studies.

The poorest 10-year survival rate was only 66% [Brown J et al., 1993]. Sample size and duration of follow-up was not reported. Callagy *et al.* reported an estimate only slightly better (73%) [Callagy GM et al., 2006].

Two studies split each of three risk groups into two, thus dividing the patients into 6 risk groups [Blamey RW et al., 2007a; Blamey RW et al., 2007b]. Results of these 2 studies could not be compared with other studies. That is because different cut offs were applied. Two cohorts were analysed in which the lowest-risk patients were defined as those with NPI≤ 2.4. The cohort with longer follow-up data gave 10-year survival rate of 88%. The corresponding rate for the other cohort was 96%.

Survival rate in the lowest risk groups at 10-years varied from 66% [Brown J et al., 1993] to 88% [Kollias J et al., 1999; D'Eredita' G et al., 2001]. However, even 88% might not be good enough to avoid treatments such as radiotherapy. Hence, there is need for new risk factors to be able to detect low risk patients with even better survival [Kirkegaard T and Bartlett JM, 2006].

Table 2.1: Comparison of 5-year event frees rates across studies in the subset of patients identified as being low risk by NPI

Study	Cohort size	Follow-up (years)	Lowest-risk group (L)		Intermediate-risk group (I)		Highest-risk group (H)	
			Number (%) of cases	**Event-free rate**	Number (%) of cases	Event-free rate	Number (%) of cases	Event-free rate
Haybittle (1982)	387	1- 6	64 (21%)	**88%**	169 (57%)	69%	65 (22%)	21%
Todd (1987)	387[2]+ 320	6-11.5 1.7- 6.5	192 (27%)	**88%**	381 (54%)	69%	134 (19%)	22%
Okugawa (2005)	311		97 (31%)	**96%**	142 (46%)	85%	72 (23%)	45%
Sauerbrei (1997)	603	5	163 (27%)	**82%**	440 (73%)	70%	No case	
Coradini (2001)	226	0.3- 8.17 Median 6.25		**82%***		72%*	No case	
Ring (2006)	195			**90%**		90%	No case	

[2] These 387 patients were those used to devise the NPI index

*inexact read off from graph

Table 2.2: Comparison of long-term event free rates (10 years or more) across studies in the subset of low risk patients identified as being low risk by NPI

Study	Cohort size	Follow-up (years)	Lowest-risk group (L)		Intermediate-risk group (I)		Highest-risk group (H)	
			Number (%) of cases	**Event-free rate**	Number (%) of cases	Event-free rate	Number (%) of cases	Event-free rate
Brown (1993)				**66%**		50%		34%
Balslev (1994)	9149	2.3- 13.9 median 7.1	2494 (27%)	**79%**	5245 (57%)	56%	1410 (16%)	25%
Kollias (1999)	2684		894 (33%)	**88%***	1374 (52%)	58%*	416 (15%)	17%*
Sidoni (2004)	82	Min 5	27 (33%)	**83%***	39 (48%)	60%*	16 (19%)	42%*
Eden (2004)	97			**83%***				43%*
Callagy (2006)	557	0.4- 39.4 Median 8.7	34 (6%)	**73%***	236 (42%)	60%*	287 (52%)	38%*
Lundin (2006)	2923	Median 9.5		**79%***		70%*		29%*
Lundin (2006)	25752	Median 9.7		**80%***		70%*		29%*
D'Eredita (2001)	402	11-19 median 15	110 (27%)	**88%***	198 (49%)	70%*	94 (23%)	40%*
Galea [3] (1992)	1629		470 (29%)	**80%***	879 (54%)	42%*	280 (17%)	13%*

[3] Galea *et al.* reported 15-year event-free rates

***** inexact read off from graph

2.2.3 Reflection on NPI

As summarised in Tables 2.1 and 2.2, NPI has been validated in several studies and its ability to distinguish low and high risk groups had been confirmed. Performance of the index in identification of low risk patients has already been discussed (see 2.2.2). However, there are methodological concerns about development of the index:

i) Cut points of risk groups

The decision about the choice of cut offs was unclear. To define low and high risk groups, 64 and 65 patients with the lowest and highest index value were selected respectively [Haybittle JL et al., 1982]. It was not revealed how the target sizes of these extreme groups were chosen. Furthermore, to form the low and risk group, including 64 and 65 patients, the split was specified as 3.4 and 5.4 respectively. However, it was not clear whether patients with an index value of exactly 3.4 and 5.4 belongs to the risk group below or above the cut off.

The nature of the NPI formula means that patients will get a score of 3.4 when tumour size is 2cm and grade and nodal status are 1 and 2 or 2 and 1. Two of the papers listed in Table 6.1, defined low risk group as those with NPI < 3.4 [Galea MH et al., 1992; Balslev I et al., 1994]. The rest of papers applied NPI ≤ 3.4 rule.

In my opinion, correct assignment into risk group is of importance. This is because wrong allocation of patients into risk groups might affect estimated event free rates. In particular, when number of patients at risk is not high, wrong risk group assignment might overestimate or underestimate the performance of model in terms of risk stratification.

ii) Missing data

My literature review showed that NPI was able to categorise patients into 3 diverged risk groups (see Table 2.2). In development of NPI, although the initial number of participants was 500 cases, a total of 79 cases with missing value were excluded (16% of data). The disadvantages of exclusion of cases with missing data and main statistical approaches to impute missing values are discussed in section 3.5.

Although exclusion of missing data in certain situations might not affect generalisability of results, in general it leads to imprecise results, and can lead to biased estimates [Altman DG and Bland JM., 2007].

2.3 Strategy in treating breast cancer

An important aim of clinical care is to maximise the survival but to avoid harsh treatments which are not needed, by optimising treatment selection in relation to prognosis. Therefore, the ability to identify a low risk group with minimal risk of recurrence is likely to have clinical appeal. That is because low risk patients could potentially avoid systemic treatment and its unwanted side effects.

Currently treatment selection for breast cancer is guided predominantly by patient prognosis, using classical pathological assessment of tumours to measure risk (i.e. NPI). In this section, treatments and side effects are given. The importance of integration of new risk factors which can improve risk stratification of patients is highlighted in section 2.4.

Ductal Carcinoma In Situ (DCIS) is the most common kind of pre-invasive breast carcinoma [Breastcancer.org, 2008a]. Some important risk factors for DCIS are age, early menarche, late menopause, older age at first pregnancy, positive family history, high fat diet, alcohol intake, smoking, weight, history of previous benign breast disease, and exposure to dioxins and ionising radiation [McPherson K et al., 2000].

In about 80% of all breast cancers, DCIS spreads into the breast tissue surrounding the ducts. It is then known as Invasive Ductal Carcinoma (IDC) which is the main focus of this section. There are two types of treatment for invasive carcinoma: local and systemic.

2.3.1 Local treatments

Local treatments treat the tumour and the surrounding area such as chest and lymph nodes and are categorised into 2 wide groups: surgery and radiation therapy.

i) Surgery

In most cases, surgery is the first line treatment. When the cancerous area is small, only the area of breast containing the cancer will be removed (known as lumpectomy or Breast Conserving Surgery (BCS)) [Breastcancer.org, 2008a].

When it is more serious, the breast, and sometimes the lining of the chest wall muscle, and some of the lymph nodes under the arm are removed (known as Mastectomy).

Side effects of lumpectomy include temporary swelling of breast, breast tenderness, and hardness due to scar tissue in the surgical site. Side effects of mastectomy

include pain and tightness in the breast area and arms, and fluid collection and infection around the operated site [Breastcancer.org, 2008a].

ii) Radiation therapy

Radiation therapy is usually recommended after lumpectomy. In this method high-energy rays are directed at the breast, chest area, and under the arms to destroy any invasive carcinoma that might be left behind [Breastcancer.org, 2008b]. Side effects of radiation involve skin reactions such as redness, itching, burning, sore and peeling.

2.3.2 Systemic treatments

Systemic treatments travel through the body to destroy any cancer cell so as to reduce the risk of recurrence or metastasis [Breastcancer.org, 2008e]. The main systemic treatments are reviewed below.

i) Chemotherapy

This method is recommended when the carcinoma is larger than 1centimetre (cm) or has spread to the lymph nodes. When chemotherapy is given after surgery, it is called adjuvant therapy. Sometimes when the tumour is large, or cancer cells have travelled to lots of lymph nodes or other parts of the body, chemotherapy might be given before surgery (known as neoadjuvant therapy) [Breastcancer.org, 2008d].

Chemotherapy has the disadvantage that, in addition to likely cancerous cells, it might also damage healthy cells in particular bone marrow, the digestive tract, the reproductive system and hair follicles.

ii) Hormonal therapy (endocrine therapy)

Hormonal therapy might be recommended when cancer cells have hormone receptors [Web of Medicine, 2008b]. The main types of hormonal therapy which are reviewed here are tamoxifen and aromatase inhibitors [Web of Medicine, 2008b].

Tamoxifen

Tamoxifen works as an anti-estrogen and blocks estrogen from attaching to estrogen receptors at cancerous cells. It decreases the chance of recurrence of early-stage breast cancers, prevents cancer development in the unaffected breast, and slows the growth of cancer cells present in the body. Tamoxifen is usually given as treatment of pre-invasive carcinoma (along with mastectomy), as adjuvant treatment of ER+ metastatic cases, as treatment of recurrent breast cancer, and as preventative treatment of women who are at high risk of developing breast cancer [Fisher B et al., 1998; Fisher B et al., 2005].

Common side effects of tamoxifen are hot flushes, vaginal discharge, irregular menstrual periods, headache, nausea and vomiting, skin rash, fatigue, fluid retention, and weight gain [Web of Medicine, 2008b]. It might also increase risk of endometrial cancer, blood clots in lung, and ovarian cysts [Web of Medicine, 2008a]. Venous Thromboembolic Events (VTEs) were increased to two-fold and endometrial cancer was increased more than two-fold in females receiving tamoxifen [Houssami N et al., 2006].

Aromatase inhibitors

This medicine blocks the effect of an enzyme that produces estrogen. Aromatase inhibitors delay the progression of breast cancer longer than tamoxifen. Side effects of these treatments are nausea, fluid retention, weight gain, and headache.

Aromatase inhibitors reduce bone density and increase fracture rates relative to tamoxifen, but they have fewer of the other side effects [Houssami N et al., 2006].

2.4 Can additional biological predictors improve risk prediction?

There is clear evidence that breast cancer is a heterogeneous disease which includes different subtypes. As an example, Estrogen and Progesterone hormonal Receptors (ER and PR), which promote growth of cancer cells, are present in nearly two thirds of breast cancer specimens [Martin M, 2006]. To evaluate the benefit from tamoxifen treatment, a series of 228 patients with median follow up of 5.8 years were analysed [Colomer R et al., 2005]. Estimated 3 and 6-year Disease Free Survival (DFS), which are summarised in Table 2.3, indicates noticeable survival difference between (ER+, PR+) and (ER-, PR-) cases.

Table 2.3: Estimated Disease Free Survival (DFS) rates in Colomer et al. study (2005)

Group	3-year DFS	6-year DFS
ER+ PR+	90%	85%
ER- PR+	82%	74%
ER+ PR-	77%	72%
ER- PR-	76%	72%

Furthermore, HER2 is a gene that controls the growth, the division, and the repair of cells. HER2 gene alteration is present in nearly 20% of tumours [Martin M, 2006]. HER2+ indicates that a protein is overproduced and therefore cells grow rapidly creating the cancer [Stephan P, 2008].

The increasing knowledge of biology, and subsequent understanding of the underlying biology of breast cancers, challenges current management of patients in which molecular difference between patients are not taken into account (i.e. NPI) [Kirkegaard T and Bartlett JM, 2006].

Molecular differences support treating different molecular sub-types based on their biology and pathology rather than pathology alone. That is because different molecular types have the potential to respond to different treatment [Kirkegaard T and Bartlett JM, 2006]. Therefore, there is a need to identify novel predictive markers and to optimise treatment selection to ensure that patients receive a treatment to which they are most likely to respond.

The NPI model has already been superseded, to some extent, by ‘Adjuvant Online’ a tool which integrates novel risk factors, for example ER and HER2 expression, with clinical trial data to select appropriate therapies for patients. The ‘Adjuvant Online’ model was developed while reviewing effectiveness of adjuvant therapy with that of tamoxifen (with or without chemotherapy) [Ravdin P, 2005] and is used across Europe and US.

It has been commented that over the next 3 to 5 years biomarkers will be incorporated as part of clinical diagnostic decision making [Faratian D and Bartlett JM, 2008]. However, many of the biomarkers measured do not yet have a clearly characterised role.

Over the past few years, the biological collaborators in this PhD research, have explored carefully the role of a large number of candidate predictive biomarkers in a selected cohort of tamoxifen treated ER+ breast cancers [Kirkegaard T et al., 2005; Cannings E et al., 2007; Kirkegaard T et al., 2007; Tovey SM et al., 2005; McGlynn LM et al., 2009]. A summary of the main results with emphasis on biomarkers which predict the outcome are presented here and summarised in Table 2.4. In each study, the focus of the authors was on a set of biomarkers which are located in the same pathway based on cancer progression. Primary outcomes studied were Recurrence Free while on Tamoxifen treatment (RFoT), Recurrence Free Survival (RFS), and Overall Survival (OS). Furthermore, a dichotomised version of biomarkers were modelled (Table 2.4) and for each univariate test patients with missing data on that variable were excluded.

Association of AKT family biomarkers with OS

Research was undertaken to investigate whether any of the 10 biomarkers grouped into the AKT family can predict OS [Kirkegaard T et al., 2005]. Authors found that, based on univariate analyses, high values of cytoplasmic Akt2 expression (n=392, P-values=0.01) and low values of cytoplasmic Pakt2 expression (n=392, P-value=0.04) were associated with higher survival. These two biomarkers were then combined to

create 3 risk groups as given below. Patients with missing value on either of these two biomarkers were excluded. The authors concluded that there is potential that tumour profiling might improve patient selection for endocrine therapies.

- Low risk group: High cytoplasmic Akt2 and low Pakt2 (n=95)
- Intermediate risk group: Both high or low (n= 185)
- High risk group: Low cytoplasmic Akt2 and high Pakt2 (n=99)

Association of BAD family biomarkers with RFS

The association between 5 biomarkers from the BAD family and RFS was assessed in univariate analysis [Cannings E et al., 2007]. Patients with high cytoplasmic Bad expression values (n=182) had better survival than those with low values (n=175), differently at a marginal P-value of 0.049.

Association between HER family biomarkers and RFoT

The association between HER1 to 4 and RFoT was assessed [Tovey SM et al., 2005]. These biomarkers were dichotomised as below. HER1 positive (HER1+) patients were those with any membranous HER1 staining (6 out of 393). For HER2, HER3, and HER4, patients were considered as being positive when at least 10% of tumour cells scored as being moderately positive (51 out of 397 for HER2, 56 of 353 for HER3, and 46 of 341 for HER4).

Plotting survival curves, HER2+ and HER3+ patients exhibited poorer survival. Furthermore, patients who were positive on any of HER1 to 3 (HER1-3+), in comparison with the remainder (98 versus 251), exhibited poorer survival.

Combination of AIB1 with HER family biomarkers to predict RFoT, RFS and OS

While no association between AIB1 and outcomes studied was seen, HER2+ patients with high level of AIB1 (n=20) have lower RFoT and OS curve than the remainder of patients (P=0.04) [Kirkegaard T et al., 2007]. A similar conclusion was drawn when AIB1 was combined with HER3. In addition, overexpression of AIB1 in HER1-3+ patients was associated with a poorer RFoT, RFS, and OS.

Association between RAS and MAPK family biomarkers and outcome (RFS and OS)

The association between the RAS and the MAPK biomarkers and cancer progression was assessed [McGlynn LM et al., 2009]. Plotting survival curves, high expression of 2 biomarkers in the MAPK family (cytoplasmic and nuclear pRaf338) were associated with worst RFS and OS. No association between the RAS family biomarkers and the rest of the biomarkers formed the MAPK family and outcomes were seen.

Table 2.4: Summary of main results of some of papers published using the data set available for this thesis

Family	Primary outcome	Split applied	Univariately informative biomarkers
AKT	OS	Median	Expressions of cytoplasmic Akt2 and Pakt2
BAD	RFS	Median	Expression of cytoplasmic Bad
HER	RFoT	See text	HER2 and HER3
RAS	RFS and OS	Top quartile	----
MAPK	RFS and OS	Top quartile	Expressions of cytoplasmic and nuclear pRaf338

2.5 Overview

Results presented indicated that the NPI was not able to identify a subset of low risk patients with very good prognosis. On the other hand, informativeness of new biomarkers has been confirmed. Therefore, there is scope to investigate whether incorporation of both biological and clinical variables improves risk prediction.

Chapter 3 LITERATURE REVIEW OF STATISTICAL METHODS FOR PROGNOSTIC MODELLING

3.1 Introduction

Multifactorial regression models are frequently used in medicine to develop prediction tools. The Framingham Coronary Heart Disease Risk Score (FCHDRS) is an example of a widely used risk score. Coronary Heart Disease (CHD) is a major cause of death and disability. The Framingham Heart Study is started in 1948 by recruiting a cohort of 5209 participants without symptoms of cardiovascular disease or heart attack [National Heart Long and Blood institute, 2009]. Since 1948, different cohorts were added to the project. In nearly 60 years, many major discoveries have been produced that enhanced the understanding of the development and progression of heart disease, stroke, and other cardiovascular disease.

The FCHDRS is a prediction tool which was developed to enable clinicians to estimate risk of CHD events for individual patients. Applying multivariate Cox PH model, classic risk factors which contribute to this model are sex, age, blood pressure, total cholesterol, low density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C), smoking behaviour, and diabetes status [Wilson PW et al., 1998].

If in development of model, one ignores model assumptions and limitations the models obtained might give false and misleading results, or might not be generalisable [Concato J et al., 1993; Wyatt JC and Altman DG, 1995].

Regression risk modelling techniques perform best when there are relatively large numbers of events and complete data for all variables [Peduzzi P et al., 1995]. However, the number of events in the cohort to be examined in this thesis is not large. Furthermore, by their nature biomarker variables are prone to missing values, and it tends to be the case that the distribution of biomarker expression is positively skewed. Therefore, the main practical challenges in prognostic modelling are to:

- Reduce the number of variables being offered to a model
- Select the appropriate form of association for variables
- Deal with the missing data
- Assess the internal validity (reproducibility) of the final model

This chapter reviews common methods proposed in the statistical literature to deal with the issues listed above. Each problem is discussed separately. The material presented here does not represent all possible solutions to these problems but it addresses those methods most frequently applied in the literature.

Review of statistical methods in the literature is difficult because useful methodological keywords are not common and any less specific keyword search might find far too many publications, many of these of scant relevance. The search strategy I used for this literature review was therefore as follows. First of all, I read some key textbooks in the field of survival analyses, as listed below. Furthermore, a reference textbook for analysis of missing data is 'Analysis of Incomplete Multivariate Data' [Schafer JL, 1997].

1. Modelling Survival Data in Medical Research [Collett D, 2003]
2. Regression modelling strategies with application to linear models, logistic regression, and survival analysis [Harrell FE, 2001]
3. Multivariable Model Building A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables [Royston P and Sauerbrei W, 2008]
4. Survival Analysis: Techniques for Censored and Truncated Data [Klein JP and Moeschberger ML, 2003]
5. Modeling Survival Data: Extending the Cox Model [Therneau TM and Grambsch PM, 2000]

6. Survival Analysis Using S: Analysis of Time-to-Event Data [Tableman M and Kim JS, 2003]
7. Analysis of Incomplete Multivariate Data [Schafer JL, 1997]

Important papers cited in the books listed above were then read. When necessary, there was follow-up reading of papers cited in these papers. Papers found to be important were followed up by searching in Pubmed to check for other papers by the same authors.

Regarding missing value and imputation methods, literature review using the key words 'multiple imputation' and 'missing data' was undertaken in Pubmed. Furthermore, some technical papers and documents were found in websites such as http://multiple-imputation.com/ and http://missingdata.org.uk/.

To avoid an over long chapter, methods are reviewed with a focus on the techniques which will be used in this thesis. Furthermore, the emphasis is on methodological issues rather than mathematical details. The remainder of this literature review is structured as follows:

Section 3.2: Types of prognostic study

Section 3.3: Modelling with many variables

Section 3.4: Exploration of functional form of association

Section 3.5: Imputation of missing data

Section 3.6: Assessment of internal validity

Examples of the application of the methods reviewed are also given. These will be distinguished from the main text and methodological review by presenting the examples in boxes.

3.2 Prognostic studies

In general, prognosis studies are divided into 2 main categories: outcome prediction and explanatory studies [Hayden JA et al., 2008]. In outcome prediction studies, the aim is to combine variables which are associated with outcome in order to stratify patients into risk groups. On the other hand, the particular aim of explanatory studies is to investigate the causal association between prognostic factors and the outcome of study [Hayden JA et al., 2008].

Explanatory studies comprise 3 main phases. In phase 1 or the exploratory phase, the aim is to identify the presence of a prognostic relationship between one or more of a set of explanatory variables and the outcome. In this phase, the aim is to describe the associations as best as possible and to generate questions about the biology of the disease. In phase 2 (confirmation phase) the aim is to confirm associations identified in the exploration step. Finally in phase 3 (understanding phase), the aim is to understand the prognostic pathway.

If the authors do not state the purpose of their study, it is very difficult to judge whether the aim of a published paper is outcome prediction or exploration (phase 1 of an explanatory study). This is because in either case, researches begin by

investigating whether there is association between a set of variables and an outcome. Subsequently, variables which are associated with the outcome are used to develop a multifactorial prediction model. Furthermore, in my opinion, some statistical issues such as multiple comparisons, might not be a serious problem in a phase 1 study, but are of concern in outcome prediction studies. Therefore it is important to carefully decide the aim and step of study.

3.3 Modelling with many variables

3.3.1 Background

The easiest solution to deal with many variables is to fit all variables in a multifactorial model (full model). In this case, there is no attempt to find a parsimonious model since all variables are in the model. To apply the model in future, information on all variables is required. However, it is often the case that the model prediction is applied using only data on variables in the model that contributed significantly, at 5% or 10% level, to the multifactorial model.

Alternatively, a more parsimonious model can be sought by utilising a stepwise variable selection procedure, such as Backward Elimination (B.E.) method. Application of stepwise methods has the advantage that to use the final model in practice, only information on a subset of key variables (those reached significance level in the multifactorial model) is required.

A problem with full fit or B.E. methods is that when many variables are offered to the multifactorial model or when the number of Events Per Variable (EPV) is low,

the full model might fail to converge to a solution. As an example, in a simulations study, at EPV of 5, in about 3% of samples convergence of Cox regression models did not occur [Steyerberg EW et al., 1999].

A more important issue is estimation of true regression coefficients. When EPV is low, estimated regression coefficients might be biased [Steyerberg EW et al., 1999; Peduzzi P et al., 1995]. However, in prognostic models the aim is to stratify patients into well diverged risk groups and therefore interpretation of regression coefficients might not be relevant.

Analysing a data set of 673 cases including 252 deaths, in the field of survival analysis, impact of EPV on estimation of regression coefficients was investigated [Peduzzi P et al., 1995]. Information on 6 binary and 1 ordinal variable were available, giving an initial EPV of 36. In the multifactorial model, fitting full model, all 7 variables were significant at a 0.10 significance level.

A series of simulation studies were then conducted with EPV of 25, 20, 15, 10, 5, and 2. In all scenarios all 7 predictors were analysed and sample size was changed. Relative bias was defined as $(b - \beta)/\beta$ where b and β indicates estimated and true regression coefficients.

At EPV of 10 or more, average bias was within ± 10%. When EVP was less than 10, the regression coefficient was overestimated by more than 20% for two variables and underestimated by 30% for one variable. Therefore, it has been commented that, as a rule of thumb, 10 Events Per Variable is advisable although 5 is the absolute minimum that is considered safe [Peduzzi P et al., 1995].

In another study, in the filed of binary outcome, a total of 40830 patients with acute myocardial infarction including 2851 deaths were analysed [Steyerberg EW et al., 1999]. Then random samples were drawn that included 3, 5, 10, 20, or 40 Events Per variable. Full model was fitted and then Backward Elimination (B.E.) variable selection method was applied. Amount of bias decreased when EPV increased (Table 3.1).

The nature of simulation studies conducted by Peduzzi *et al.* and Steyerberg *et al.* had some limitations. In neither of them was overfitting evaluated since no non-significant variable was included in the model. Furthermore, since sample size but not number of variables were changed the question remains whether results would be similar if the number of independent variables varies.

Table3.1: Range of relative bias at different EPV's for full model and B.E. variable selection methods in Steyerberg et al. study (1999)

EPV	Full model	B.E. variable selection method
3	-20%, 15%	14%, 206%
5	-12%, 8%	4%, 159%
10	-7.5%, 4.2%	2%, 110%

Recently a large simulation study was conducted, for both binary outcomes and time-to-event data, focusing on a primary predictor (either continuous or binary) and regarding the covariates as adjustment variables [Vittinghoff E and McCulloch CE, 2007]. Different combinations were considered: EPV (2, 4, 6, 8, 10, 12, 14, 16), sample size (128, 256, 512, 1024), value of the regression coefficient (0, log (1.5), log (2), log (4)), and models with a total of 2, 4, 8, and 16 predictors. Relative bias higher than 15% was considered as a major problem. For the Cox model with continuous predictor, when EPV varied from 2 to 4, relative bias was higher than 15% in 17.2% of analyses. The corresponding figure for EPV 5 to 9 was higher than 15% only 2%. It has been commented that, to control for effect of confounding, the EPV rule could be relaxed [Vittinghoff E and McCulloch CE, 2007]. Results should be compared with models in which weak predictors are not offered to the model.

As an alternative to traditional Cox regression model, Tree-based Survival Model (TSM) can be applied which begins with all variables in contention but avoids the problem of convergence of regression models. TSM involves successive binary partitioning, classifying subjects into smaller groups. At each step, every possible cut point for each prognostic variable will be examined to select the split which best discriminates patients, based on patients' outcome. Typically the resulting model is presented graphically as a decision tree.

While traditional multifactorial regression tools (such as Cox without interaction term), for the whole sample suppose a uniform effect of the variable, TSM can reveal

factors with different effects in different subgroups. This is biologically plausible that a variable be important for only a subset of patients. From that perspective, TSM is a complement for Cox regression model and hence has potential benefits in terms of therapeutic management [Banerjee M et al., 2004].

It has been claimed that TSM provides a readily interpretable picture, results in easier clinical decision making, and aids future studies [Harrell FE et al., 1998; Banerjee M et al., 2004; Banerjee M et al., 2004; Ciampi A et al., 1988; Ture et al., 2009].

In my opinion, a tree structure is not always simple to interpret and to be used in practice. That is because sometimes patients in the different nodes have similar survival curves and therefore amalgamation of such groups are required. This cannot be detected by simply looking at a tree but after plotting Kaplan-Meier curves and comparing event-free rates in final groups.

In TSM, there is no limit on the number of variables involved [Therneau TM and Atkinson EJ, 1997]. However, a disadvantage is that TSM might lead to unstable results due to the extensive cut point search [Hukkelhoven CW et al., 2006].

3.3.2 Methods to reduce number of variables prior to modelling

As an alternative to methods which begin with all variables (see section 3.3.1), one could apply a data reduction step prior to modelling. In this section, I will review common data-reduction approaches and their advantages and potential disadvantages. A summary of the methods reviewed is provided in Table 3.3 at the end of this section.

i) Univariate screening

The standard statistical approach is to use a pre-specified univariate screening process to select a reduced subset of variables to be offered to the multifactorial modelling stage. This reduction in the number of variables is however at the cost multiple comparisons. Furthermore, in the case that confounding is present, important predictors might be missed at the univariate screening stage [Hosmer DW and Lemeshow S, 2000; Sun GW et al., 1996].

This issue is illustrated by analysis of a hypothetical data set of 4207 patients to investigate the effect of sex on occurrence of Coronary Heart Disease (CHD). In univariate logistic regression, no association was found (P=0.20). Authors explained that physical activity is a known risk factor for CHD and therefore stratified analysis was performed. In univariate logistic regressions, significant association between sex and CHD was seen in both inactive and active groups (P-value =0.028 and 0.026 respectively). Furthermore, both variables were retained in the multifactorial logistic regression model. Relationship was detected either by controlling of confounder or by performing multifactorial regression model.

ii) Principal Component Analysis (PCA)

Use of Principal Components Analysis (PCA) allows the original variables to be replaced to a smaller number of components where these components explain the majority of variation in original variables. Each component is a linear combination of the original variables [D'Agostino RB et al., 1995]. Components derived then can be used in regression modelling [Harrell FE et al., 1985; D'Agostino RB et al., 1995; Marshall G et al., 1994].

Example:

A predictive model for patients undergoing coronary artery surgery was developed [Marshall G et al., 1994]. A total of 6317 participants including 285 events formed the sample.

PCA was applied to 33 variables [Marshall G et al., 1994]. These were reduced to 5 components which were offered to the modelling.

Application of PCA means that fewer variables are tested in the model. However, since each component is a combination of original variables, estimated Hazard Ratio's (HR) are not simple to interpret. Furthermore, to be used in practice, one has to measure all covariates (or at least all those with high contribution to components) [Harrell FE et al., 1985; Weber G et al., 2004]. In addition, components derived, despite capturing dimensions of variance in the potential explanatory variables, might not be optimally informative with respect to the outcome of study.

iii) Biologically informed data reduction method

In biologically guided approaches, external knowledge guides model building. Variable selection involves two steps [D'Agostino RB et al., 1995]. The first step is to divide the variables into substantive sets based on biological knowledge (e.g. tumour progression pathway). In the second step, a composite score (also known as sickness score) will be calculated for each family set. The resulting reduced number of composite scores derived will then be offered as intermediate predictors to the multifactorial model. This approach is inherently Bayesian since prior knowledge guides model development, although no formal Bayesian analysis would be undertaken.

Alternative approaches to calculate a composite score are to sum the variables with or without weighting. With binary or categorical data, the simplest approach is to count the number of positive characteristics (e.g. for a group of symptoms, the score is the number of symptoms presented) [Harrell FE et al., 1984].

Example:

This approach is used to address whether multiple Single Nucleotide Polymorphisms (SNP) can add value to that of traditional risk factors in prediction of Coronary Heart Disease (CHD) [Morrison AC et al., 2007]. A total of 1452 participants were genotyped for 116 SNPs. Only 22 SNPs with univariate P-value less than 0.10 were candidates for the final model. Each SNP was a categorical variable (1: risk homozygote, 0: heterozygote, -1: non-risk homozygote). SNPs were summed together. The score derived was significantly associated with CHD.

Summation of features of binary or categorical data is straightforward to implement but is not possible for continuous data. As an example, 25 variables were used to predict complete remission from treatment of 334 patients with non-Hodgkin's lymphoma [Harrell FE et al., 1985]. However, only 14 binary variables were used to define a sickness score [Harrell FE et al., 1985]. Another problem with counting of binary data variables is that this method gives equal weight to each variable involved, regardless of relative importance.

An alternative method is to give a weight for each variable where weights reflect the relative biological importance [Harrell FE et al., 1984; Marshall G et al., 1994; Marshall G et al., 1995].

Specification of weights based on key relevant biological knowledge is attractive. However, it is often the case that the biological knowledge to do this is unknown.

Furthermore, no parsimonious index can be found in the case lots of variables contributed to it. In addition, this method does not reduce number of variables that needs to be measured.

3.3.3 Research comparing data reduction techniques

Four papers compared performance of alternative data reduction techniques in terms of discrimination ability, or C-index (Table 3.2). This statistics varies from 0.50 to 1. The higher the C-index, the higher the discrimination ability is. Further detail of this statistic (C-index) is explained in section 4.4.7 (i).

Marshal *et al.* showed that the B.E. stepwise regression model had the highest discrimination ability, followed by TSM. Biologically informed approach had the poorest performance [Marshall G et al., 1994]. This might indicate that weights given to calculate the sickness score were not appropriate.

Harrell *et al.* (1984) showed that performance of logistic regression and PCA were the same and much better than tree-based method. At EPV of 2, performance of the tree-based method was superior to regression model. Furthermore, PCA showed the best performance [Harrell FE et al., 1985]. This might indicate that, at very low EPVs, regression models might not work well.

In another paper, analysing 2113 patients including 208 deaths, a total of 30 variables were used to predict artery disease (EPV=6.9) [Harrell FE et al., 1984]. Different training samples were then drawn corresponding to EPV's 3.4, 1.8, and 1. In all

scenarios, performance of stepwise logistic regression was marginally superior to PCA [Harrell FE et al., 1984]. In my opinion, components derived might not necessarily always be optimally informative to the outcome and explain poorer performance of PCA.

Using 50 variables to predict myocardial infarction in 482 patients with chest pain, stepwise logistic regression and tree-based methods were applied [Cook EF and Goldman L, 1984]. Tree based method worked better than regression model. As EPV was not reported it might not be simple to judge whether that was due to small EPV or structure of their data set.

In general logistic regression models showed higher discrimination ability than PCA and tree-based methods. However, when EPV was very low performance of tree-based methods was superior to the regression model.

Table 3.2: Discrimination ability (C-index) of standard method versus alternative methods to deal with many variables

Study	# events (variables)	EPV	C-index			
			Regression model (B.E.)	PCA	Biologically informed	Tree-based
Marshall *et al.* (1994)	285 (33)	8.6	0.74	0.70	0.67	0.72
Harrell *et al.* (1985)	102 (25)	4.1	0.67	0.68		0.56
	50 (25)	2	0.58	0.67		0.61
Harrell *et al.* (1984)	208 (30)	6.9	0.85	0.81		
	101 (30)	3.4	0.84	0.83		
	55 (30)	1.8	0.85	0.84		
	30 (30)	1	0.88	0.86		
Cook *et al.* (1984)	Not Given					

3.3.4 Summary

A summary of main approaches to deal with many variables and their features are given in Table 3.3. Each of methods reviewed has its own limitations and therefore it is not simple to judge which technique is the best method in all data sets.

Table 3.3: Commonly used approaches for dealing with many variables

Method	For future applications, information on only a reduced set of variables is required	Interpretation of results is simple	Minimal statistical work required prior to modelling	Other potential problem
Methods involving no data reduction				
Fit model offering all variables (in conjunction with B.E. variable selection method)	Yes	Yes	Yes	Might not converge Regression coefficients might be biased
Tree-Based Survival Methods	Yes	Yes	Yes	Instability
Prior reduction of number of variables				
Screen for informative variables	Yes	Yes	No	Requires multiple comparisons. Confounder effects might not be simple to check
Principal Component Analysis (PCA)	No	No	No	Components might not be informative
Define substantive sets, based on biological knowledge, and calculate an index for each set to be used in regression model	No	No	No	Inadequate knowledge Confounding between variables in different families are missed

3.4 Methods to ascertain the appropriate form for continuous variables

3.4.1 Background

The Cox regression model is frequently used to analyse follow-up data. One of the most important assumptions for this model is the linearity of effects. This means that the effect of a variable is monotonic increasing or decreasing. As an example, for age this requires that the hazard ratio between a 45 and a 50-year old must be the same as that between an 80 and 85 year-old [Therneau TM and Grambsch PM, 2000].

Yet, a recent review of 99 articles published in 2 major epidemiology journals (Journal of Clinical Epidemiology and American Journal of Epidemiology) showed that fewer than 20% of papers using multifactorial logistic regression described conformity for linearity gradient [Ottenbacher KJ et al., 2004].

In the case of laboratory measurements, there is a considerable chance that the linearity assumption might not be justified [Hastie T et al., 1992]. Furthermore, when the relationship is J or U shape, a linear risk function will be unlikely to capture the relationship in a way helpful to model fit. Therefore, when evaluating the contribution of covariates on disease course, it is of importance to establish the correct functional form of any continuous covariates [Hastie T et al., 1992].

In a recent study, the performance of different statistical techniques in terms of estimation of functional form was compared by carrying out a simulation study

[Hollander N and Schumacher M, 2006]. I firstly shall review methods available to ascertain the appropriate form for continuous variables and after that, I will review the results of the simulation study conducted by Hollander *et al.*

In this chapter, I classify the nature of association as either 'linear or polynomial', or 'threshold' effects (where values at or above a specified level predicts outcome).

3.4.2 Linear or polynomial forms

The most common approaches used to estimate linear or polynomial effects are polynomial regression models, General Additive Models (GAM), and Fractional Polynomial (FP). These methods are reviewed and advantages and disadvantages are described.

i) Polynomial regression modelling

One extension to a standard Cox model is to allow for a polynomial relationship by adding a polynomial term such as quadratic to the model [Therneau TM and Grambsch PM, 2000; Harrell FE, 2001].

Example:

Analysing 477 participants, quadratic association between diet factors (such as protein, oleic acid, cholesterol, and percentage of calories from fat and carbohydrates) and breast cancer was reported [Goodwin PJ et al., 2003].

ii) Generalised Additive Modelling (GAM)

With polynomial regression the fit is global, so a localised pattern might be obscured [Hastie T et al., 1992]. GAM models (regression splines and smoothing splines) are flexible techniques that allow the detection of polynomial associations that vary along the range of the covariate. Spline functions are piecewise polynomials within pre-determined intervals of the covariate. The general form of GAM model is [Hastie T et al., 1992]:

$$h_i(t \mid x_1, x_2, \ldots x_p) = h_0(t) \exp\{\sum_{j=1}^{p} s_j(x_j)\}$$

Here, the fitted curve only depends at each point on observations at that point, and within a pre-specified neighbourhood. In a regression spline model, the researcher decides the form of the function.

Regression splines such as Restricted Cubic Splines (RCS) allow detection of highly curved associations [Harrell FE, 2001]. RCS uses a linear fit in the tails (before the first and after the last knot, where knots are the points at which the X axis is divided into intervals, and which are also parameters of the model that must be pre-specified).

Regarding the position of knots, a well-accepted approach is to put the knots at quartiles. This makes a trade-off between flexibility and loss of precision due to overfitting a small sample. However, it has been reported that when the sample size is large, use of five knots is a good choice [Stone JC, 1986].

Examples:

1) GAM was used by in a series of 265 post-menopausal breast cancer patients with 6.86 year median follow-up [Hastie T et al., 1992]. A non-linear effect of number of nodes, age and body mass index on Disease Free Survival (DFS) was reported. Age showed an inverse U-shape association where the death hazard remained constant for those aged 50 to 60 years old. Furthermore the curvature association for number of nodes and Body Mass Index (BMI) suggested a threshold effect at 16 for number of nodes examined, and at 32 for BMI [Hastie T et al., 1992].

2) Modelling the effect of age in 83804 breast cancer patients, a biphasic association with all cause mortality was found, suggesting two age components: a linear component (corresponds to a natural increase of mortality with each year of age) and a quasi-quadratic component reflecting an increased risk of mortality for patients older than 50 years old [Tai P et al., 2005].

iii) Fractional Polynomial modelling (FP)

Fractional Polynomial (FP) modelling is a powerful tool to detect non-linear associations [Royston P and Altman DG, 1994]. Prior to availability of FP modelling, researchers would usually apply logarithmic transformation to the skewed data. Other commonly used transformations are the square, cubic, or reciprocal. All of these transformations are embedded in FP method. FP selects the optimum power transformation by testing a range of values (see technical details in section 4.4.2).

Example:

FP method was used to detect the best functional form for age and progesterone receptor in a series of 686 node-positive breast cancer patients [Sauerbrei W et al., 2006]. It was revealed that patients aged less than 40 had a markedly increased risk of recurrence, followed by a fairly constant plateau for those aged 40 to 55, with a slight increase again after 55. Furthermore, a logarithmic transformation was proposed for progesterone receptor.

While FP is flexible in its ability to detect polynomial effects, extreme values in the covariate might result in unstable transformations [Royston P and Sauerbrei W, 2007]. As an example, analysing 252 men, a polynomial relationship between percentage body fat and abdominal circumference was seen [Royston P and Sauerbrei W, 2007]. Since extreme values have high leverage, one case with very high value was excluded from analysis. Linear model was then adequate (P=0.70 for polynomial versus linear).

3.4.3 Threshold forms

In medical applications researchers often dichotomise continuous covariates prior to modelling analyses. From a statistical point of view, dichotomisation eliminates the need for the linearity assumption, makes data summarisation more efficient, and allows for simple interpretation of results [Williams BA et al., 2006]. In the regression setting, for instance, interpretation of the impact of a binary covariate on

outcome is easier than that for a change of 1 unit in a continuous covariate [Therneau TM and Grambsch PM, 2000; Harrell FE, 2001].

Furthermore, it has been claimed that, from the clinical point of view, binary covariates might be preferred because they offer a simple risk classification into high versus low, assist in making treatment recommendations, and in setting diagnostic criteria [Mazumdar M and Glassman JR, 2000; Williams BA et al., 2006].

On the other hand, dichotomisation can result in the loss of information and power, if a linear rather than threshold association pertains, and non-linear relationships such as U-shape associations will not be detected [Altman DG and Royston P, 2006; MacCallum RC et al., 2002].

This issue has been illustrated in an analysis of 207 patients with primary biliary cirrhosis [Royston P et al., 2006]. The association between 2 continuous and 2 binary variables, and treatment was evaluated. Different multifactorial models were developed in which continuous variables were modelled in continuous and binary form. The model in which continuous data were treated as being continuous had highest discrimination ability and model goodness of fit [Royston P et al., 2006].

It has been emphasized that dichotomisation is appropriate only when a threshold effect value truly exists. That is, if we can assume some binary split of the continuous covariate creates two relatively distinct but homogeneous groups with respect to a particular outcome [Abdolell M et al., 2002]. Due to the widespread use

of dichotomised variables in medical literature, common methods to dichotomise continuous variables with their advantages and disadvantages are reviewed below and summarised in Table 3.6.

i) Dichotomisation based on biological knowledge

Dichotomisation based on biological evidence is the most attractive method. As an example, researchers sometimes dichotomise age of women at 50 years to be a surrogate for approximate menopausal status. However, for the majority of variables the biological knowledge needed is not available.

ii) Dichotomisation at a pre-specified point

Another method commonly used is to categorise the covariate at a pre-determined split such as the median [Linderholm B et al., 2000]. In this way an equal proportion of patients (50%) are assigned to each group.

However, dichotomisation at median leads to different threshold values from one study to another, and creates difficulties in comparing findings across different studies [MacCallum RC et al., 2002]. As an example, in a meta analysis of eleven studies on the role of cathespin D on Disease Free Survival (DFS) of breast cancer patients, the cut points used to define high/low cathespin D concentration ranged from 20 to 78 [Ferrandina G et al., 1997].

iii) Dichotomisation based on Martingale residuals

The Martingale residual for an individual is the difference between the observed event status and the expected value predicted with the Cox model. Plotting Martingale residuals against the value of a covariate provides pictorial evidence to investigate a threshold effect [Bradburn MJ et al., 2003]. If there is a threshold effect, the plot should display an S-shape curve. A linear plot implies adequacy of linear fit [Klein JP and Moeschberger ML, 2003].

Example:

Martingale residuals were plotted to study the functional form for number of involved and uninvolved axillary nodes in early breast cancer mortality [Vinh-Hung V et al., 2003]. For the number of uninvolved nodes, death hazard decreased but stabilized beyond 5-10 uninvolved nodes. For the number of involved nodes, no clear cut off was apparent as hazard mortality continued to increase with each involved node.

iv) Minimum P-value method and 'Classification And Regression Trees' (CART)

When there is no biologic evidence or priori information regarding the underlying relationship between the covariate and the outcome, it is possible to seek the cut point which gives us the largest difference between individual outcomes in the resulting two groups [Klein JP and Moeschberger ML, 2003].

In the minimum P-value approach, after a systematic search across all possible values, the value chosen as the cut point will be that with the smallest corresponding P-value in a Log-Rank test, when comparing the survival curve of two groups formed [Lausen B and Schumacher M, 1992]. However, Heinzl *et al.* warned that 'without any biological or clinical indications for the actual existence of a cut point or dangerous segment even the correct application of the minimum P-value approach has to be considered methodologically inferior' [Heinzl H and Tempfer C, 2001].

Example:

S-Phase Fraction (SPF) refers to the proportion of cells in the S phase of the cell cycle which reflects rate of tumour proliferation. Analysing 169 node-negative breast cancer patients, minimum P-value method was used to find a cut point for S-phase fraction. The outcome of study was Recurrence Free Survival (RFS). A split at 10% was identified [O'Reilly SM et al., 1990].

Classification And Regression Tree (CART) is simply the extension of minimum P-value method to multiple covariates using a tree structure. In this approach, after creation of two groups, the process will be continued in a branching system to categorise patients on the basis of further covariates or a previously used covariate at a new cut point. These approaches (minimum P-value and CART) require multiple testing and hence might give unstable cut points [Clark TG et al., 2003]. Alternative methods to improve the stability and to correct for multiple testing are reviewed below.

Improving stability of minimum P-value method

The evidence of stability of a cut point model is required by performing with graphical (minimum P-value graph) and numerical methods (bootstrap study) [Mazumdar M and Glassman JR, 2000]. In addition, to avoid groups with very small/ high number of patients, it has been recommended not to apply any split at the outer 20% of the covariate distribution [Lausen B and Schumacher M, 1992; Altman DG et al., 1994].

A minimum P-value graph plots all cut point values of covariate against corresponding P-values to assess whether any other cut off(s) exists with P-value similar to that of minimum P-value [Dannegger F, 2000]. Furthermore, to reduce instability, the median of optimum splits across bootstrap samples can be used as the split. However, in the case that competing cut offs are far from each other, the use of a modal statistic is preferable [Dannegger F, 2000].

Multiple testing

Multiple testing is a regrettable consequence of minimum P-value method. Use of this method might result in a type one error as high as 40% [Altman DG, 1998]. This rate might be inflated to 50% if examining 50 cut points [Hilsenbeck SG et al., 1992]. Therefore, a cut point P-value should be adjusted to reflect multiple testing, and to reach a decision regarding whether or not to adopt the cut point.

Alternative methods to deal with multiple testing are to apply two-fold cross-validation, to perform sample-split techniques, and to correct the P-value obtained [Hilsenbeck SG and Clark GM, 1996]. These methods are reviewed below.

In a two-fold cross-validation approach the data is divided into two equally sized subsets. Minimum P-value method is applied separately in each subset to find the optimal cut points (say C1 for first subset, C2 for second subset). Cut points derived are applied to the other subset. The subgroups of patients with low values for the covariate is a combination of the below cut point patients in each subset. High risk patients are defined in a similar way. The P-value of the covariate is estimated using a Log-Rank or Cox model [Mazumdar M et al., 2003]. Simulation studies show that the type one error for this method is approximately correct [Faraggi D and Simon R, 1996].

In the sample-split method, the data will be divided into training and test samples. The optimal cut point derived in the training set will be applied in the test set to find the correct P-value.

Example:

The danger of using an optimal cut point without adjustment of P-value has been stressed [Hollander N and Schumacher M, 2001]. A series of 686 node-positive breast cancer patients was divided into two equally sized samples (training and test samples). A Minimum P-value method was applied to the training sample. An optimal split for age at 43 years old was proposed in the training set (unadjusted P-value= 0.02). On the other hand, if the adjusted P-value had been calculated for the training sample, it would have been 0.31, far from significant and preventing a misleading impression of association with age younger or older than 43.

Applying this cut point to the test sample gave a P-value of 0.23. Furthermore, application of this split to another independent sample (n=139) resulted in P-value of 0.38 [Hollander N and Schumacher M, 2001].

It might be that neither two-fold cross-validation nor two-sample statistics are feasible when sample size and number of events is low. On the other hand, P-value correction methods and nomograms have been developed to correct a P-value obtained [Hilsenbeck SG and Clark GM, 1996]. If ε_{low} and ε_{high} show the proportion of the observations at the bottom and top of the highest cut point value considered, derivations below were proposed for $\varepsilon_{low} = \varepsilon_{high} = 0.05$ and $\varepsilon_{low} = \varepsilon_{high} = 0.10$ [Altman DG et al., 1994].

$$P_{alt5} = -3.13 P_{\min}(1 + 1.65 Ln(P_{\min}))$$

$$P_{alt10} = -1.63 P_{\min}(1 + 2.35 Ln(P_{\min}))$$

3.4.4 Comparison of methods to estimate form of association

A comprehensive simulation study was conducted to compare the ability of alternative statistical methods to estimate the correct form of risk function [Hollander N and Schumacher M, 2006]. Techniques compared were

- Linear Cox model
- Polynomial regression model (linear plus quadratic terms)
- Generalised Additive Models (Restricted Cubic Splines (RCS))
- Fractional Polynomials (FP)
- Categorisation at fixed points
- Minimum P-value method and its extension CART

A continuous variable X which was uniformly distributed on the interval [1, 2] was simulated. The survival time T from an exponential distribution was generated by using the transformation $T = (-1/\lambda)\log(U)$, where U was uniformly distributed on [0, 1] and $h_i(t) = h_0(t)\exp(g(x,\beta))$ with baseline hazard $h_0(t) = 1$ was chosen according to one out of 4 different risk functions g :

- Null model: $g(x,\beta) = 0$
- Cut-point model: $g(x,\beta) = \beta * I_{\{x>\mu\}}$ with μ=1.5, β=0.5
- Linear effect model: $g(x,\beta) = \beta x$ with β=0.5
- V-type effect model: $g(x,\beta) = 2\beta|x-\mu|$ with μ=1.5, β=0.5

Simulation study was performed with 100 observations and 1000 replications. Methods were compared in terms of Mean Absolute Error (MAE) which was the absolute difference between the standardised estimated and simulated risk functions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left| \hat{h}(x_i, \beta) - g(x_i, \beta) \right|$$

The 95th percentile of the empirical distribution of MAE are summarised in Table 3.4. When there was no real association (null model) and when association was linear, FP produced a MAE of 0.19. This rate slightly increased to 0.22 and 0.23 when the true form was cut point or V-shape. Comparing methods which are particularly designed to detect polynomial associations (polynomial regression, GAM (RCS), and FP), FP and GAM produced lowest and highest MAE respectively (Table 3.4).

Between methods which are devised to detect threshold effects (dichotomisation at fixed point, Minimum P-value, and CART), dichotomisation at a fixed point showed the lowest MAE. Furthermore, when the P-value obtained was not corrected, Minimum P-value and CART methods had a very poor performance. The MAE was higher than 0.35 for all simulated forms. Type one error increased to more than 0.40 (data not shown). Therefore, use of CART and the minimum P-value method without adjustment of P-values obtained has been criticised, on account of its tendency for overfitting [Hollander N and Schumacher M, 2006]. These techniques, when there was no association at all, detected a false effect in 43% of replications (data not shown).

However, considerable reduction in MAE was seen once P-values obtained were corrected for multiple comparisons undertaken.

In general, FP produced the lowest MAE, and was the only method which holds the correct type one error (data not shown) [Hollander N and Schumacher M, 2006]. Performance of polynomial regression and dichotomisation at a fixed value was approximately the same and marginally better than GAM.

Some of the limitations of this study are that only univariate analyses were undertaken, and in simulated threshold and v-shape associations only a moderate change in risk was simulated. Furthermore, no polynomial form was simulated.

Table 3.4: The 95th percentile of the empirical distribution of the difference between the true and estimated form of association across techniques applied in the simulation study by Hollander *et al.* (2006)

True Association	Techniques used to ascertain form					
	Polynomial regression	GAM (RCS)	FP	Dichotomisation at fixed point	Minimum P-value (and CART) (without P-value correction)	Minimum P-value (with P-value correction)
Null model	0.21	0.25	0.19	0.21	0.36	0.23
Linear model	0.21	0.24	0.19	0.19	0.37	0.27
Cut-point model	0.23	0.25	0.22	0.23	0.37	0.29
V-shape model	0.23	0.26	0.23	0.22	0.35	0.27

3.4.5 Summary

Advantages and disadvantages of methods reviewed to estimate linear or polynomial effects are summarised in Table 3.5. FP, in comparison with other methods, makes the most use of data and optimises power transformation. This method maintains correct type one error. Furthermore, in comparison with GAM, results are easier to communicate [Sauerbrei W et al., 2007].

Table 3.5: Advantages and disadvantages of methods to detect polynomial effects

Method	Advantage(s)	Disadvantage(s)
Polynomial regression modelling	Easily understood Simple to implement	Selection of polynomial degree is subjective Prevent selection of optimum power
Generalised Additive Modelling (GAM)	Able to detect highly curved associations	Results difficult to communicate Might slightly increase type one error Curves can be difficult to interpret
Fractional polynomial (FP) modelling	Makes the most use of data while maintains correct type one error for each variable Good approximate for threshold and V-shape effects	Powers selected will be sensitive to extreme values so confirmation of stability of transformation is required

Dichotomisation of continuous data is a contentious issue. That is because although it has biological appeal, it might lead to waste of information. However, in exploratory studies of biological variables, it can be worthwhile to explore the data to investigate this sort of association. This helps to extract more information from the data which might be useful in understanding of biological mechanisms of disease. Main advantages and disadvantages of methods reviewed are summarised in Table 3.6.

Table 3.6: Advantages and disadvantages of methods to detect threshold effects

Method	Advantage(s)	Disadvantage(s)
Use of biological knowledge	Simple and attractive	Knowledge in most cases inadequate
Use of median (pre-specified)	Avoids accusation of data dredging Provides balanced distribution of data in to two groups	Findings across studies might not be comparable
Martingale residuals	A simple and informative graphical tool	Choice of split is subjective
Minimum P-value method	Optimises place of split Useful in exploratory studies Might find false associations	Requires p-value correction and check for stability P-value can be corrected using Altman formulas or nomograms available Stability can be improved using median or mode of optimal split across bootstrap samples

3.5 Handling missing data

3.5.1 Background

In a recent review of 100 papers reporting survival analysis, published in 2002, only 15 papers involved no missing data [Burton A and Altman DG, 2004]. A total of 81 papers had data with missing covariates and 4 papers did not provide sufficient information to determine whether there was missing data [Burton A and Altman DG, 2004]. In this section, mechanisms for missing data and main approaches to deal with missing data are reviewed.

3.5.2 Missing data mechanisms

Since the mechanism of missing data is an important issue in comparison of methods used frequently to deal with missing data, possible mechanisms are reviewed. Missing data are divided into three types: Missing Completely At Random (MCAR), Missing Not At Random (MNAR), and Missing At Random (MAR) [Altman DG and Bland JM., 2007].

Missing data are categorised as MCAR when subjects with missing data are a random sample of data [Donders AR et al., 2006]. For example, MCAR occurs when a blood sample tube is broken or when a questionnaire is accidentally lost.

Categorisation of missing data as MNAR applies when the probability that an observation is missing is related to unobserved information, such as the actual

(missing) value of the variable [Rubin DB, 1976]. This can happen, for example, when a patient is so sick that a medical procedure cannot be applied to measure a study variable.

However, missing data are usually neither MCAR nor MNAR but MAR [Schafer JL, 1997]. MAR applies when the probability that an observation is missing is related to other observed patient characteristics. For example, in multicentre studies some centres might not collect data on a particular variable [Altman DG and Bland JM., 2007].

3.5.3 Approaches to deal with missing data

The main methods to deal with missing data are to exclude patients with missing data on any variable (Complete-Case analysis), to replace them with a fixed value such as mean or median, to apply the Expectation Maximum (EM) algorithm, or to use a multiple imputation technique. These methods with their advantages and disadvantages are summarised in Table 3.7.

i) Complete-Case (C-C) analysis

Complete-case analysis, which is the exclusion of cases with missing data on any of variables under study, is the simplest way to proceed. A review of 100 papers found that complete-case analysis was the method most frequently used [Burton A and Altman DG, 2004].

Under the MCAR assumption, subjects with complete data are a random sample of data and therefore if a small proportion of entire data set, less than 5%, is missing, case deletion is a reasonable approach [Fairclough DL, 2004]. However, when missing rate is high, exclusion of missing data will diminish precision of estimates.

To explain effect of exclusion of missing data on precision of estimates, a total of 2800 cases with complete data on Pre-Hospital Index (PHI) were analysed [Joseph L et al., 2004]. PHI is used in the evaluation of trauma care. Data were then omitted randomly under MCAR mechanism.

The 95% C.I. for mean of PHI, in original and reduced data (N=2800 versus 956) were (3.40, 3.96) and (3.15, 4.11) respectively. Exclusion of cases with missing data resulted in the loss of precision and a wider confidence interval.

In another study, a cohort of 300 subjects was simulated using 500 computer replications. Simulated data sets consisted of a dichotomous outcome and 3 variables: a binary exposure, a continuous confounder, and a binary confounder. True OR was 3 for binary exposure variable. Missing data were generated by MCAR mechanisms at attrition (missing) rates 10%, 25%, and 40% [Kristman VL et al., 2005].

Estimated OR's (S.D.) at various attrition rates were 2.98 (1.85) at 10% attrition, 2.81 (2.54) at 25% attrition and 3.05 (2.84) at 40% attrition. The higher the missing rate, the larger the estimated SD was.

ii) Replacement of missing values with mean or median of observed values

A method frequently used in the literature is to substitute missing data by the mean or, in the case of skewed data, by median of observed values. This might artificially reduce the variance and affect the strength of relationships with other variables [Donner A, 1982; Croy CD and Novins DK, 2005].

Using a data set of a study on substance use among American Indian adolescents, out of 209 patients, only 161 participants (76%) with available data were analysed [Novins DK et al., 2001]. This gave mean age at first use of 14.66 (SD 2.5). corresponding figures after substitution of missing data by mean was 14.59 (SD 2.3) respectively [Croy CD and Novins DK, 2005].

Furthermore, in case-control studies, replacement of missing data with a fixed value increases the overlap between cases and controls and will hence tend to underestimate the true association. This has been demonstrated by simulating 1000 samples of 500 subjects, consisting of equal numbers of diseased and non-diseased subjects and a continuous diagnostic test [Donders AR et al., 2006]. The true Odds Ratio (OR) between diagnostic test and disease status was 2.7.

Following a MCAR mechanism, the diagnostic test values for 20% of diseased and 20% of non-diseased subjects were omitted. When these missing data were replaced with overall mean, the estimated OR was reduced by 36% to 1.73.

iii) Expectation Maximum (EM) algorithm

Expectation-maximum (EM) algorithm is a likelihood based method. The philosophy behind the EM method is that if values for missing data are known, then estimation of model parameters is straightforward. Similarly, if the parameters of the model are known, then it is possible to replace missing data with unbiased values. EM works by combining these two steps. It first estimates the parameters on the basis of the data available, and then estimates the missing data on the basis of those parameters, with these two steps continuing iteratively.

In the case of skewed data, to avoid out of range values, a log-transformation of the predictors might be applied before imputation of missing data [Schafer JL, 1997]. However, after the imputation process an anti-log transformation has to be employed to bring values back to the original scale. For binary and categorical data, a rounding approach to the nearest possible value, might work well in practice [Schafer JL, 1997].

The EM algorithm preserves the characteristics of data (such as mean, correlation, and variance). This issue is illustrated analysing a sample of 492 patients, from a longitudinal study on the stress and health of elderly adults [Musil CM et al., 2002]. Mean for all cases (n=492) was 7.34 (SD 7.28). Data on a single variable for 96 cases (20%) was dropped out under MAR. Replacement of missing data by mean underestimated true SD, in that the estimated mean was 6.38 (SD 6.12). Applying EM algorithms gave estimates of 6.79 (SD 6.23) and correlation between variable under focus and rest of variables was fairly similar to that of original data.

iv) Multivariable Imputation via Chain Equations (MICE)

The MICE method is a powerful tool to tackle the missing values. In this method the aim is to preserve data features (mean, variance, correlation) while taking into account the uncertainty regarding what the unknown missing value would have been if not missing.

The MICE method replaces each missing value by multiple imputed values, resulting in multiply imputed data sets. This is a probabilistic approach which reflects the uncertainty about the true values of the missing data [Schafer JL, 1997]. It has been shown that 3 imputed data sets are sufficient when 20 per cent of values are missing. In general, there might be little or no practical benefit to create more than 5 to 10 imputed data set [Schafer JL, 1999].

It has been suggested that, for the best imputation the outcome variable should be included in the imputation model [Moons KG et al., 2006]. Data from 364 subjects in a prospective diagnostic study on pulmonary embolism were used. Five predictors without missing value were selected and estimated regression coefficients were considered as true values. Between 10% and 15% of values were omitted under MCAR and MAR mechanisms. MAR mechanism was created by assigning missing value to predictors that was related to other predictors and outcome. The MICE method was then applied to impute missing data. This was done with and without use of an outcome variable in the imputation model. Results were presented graphically. It has been seen that, the amount of bias in the estimated regression coefficients and intercept were much lower when outcome was included in the imputation model.

However, presence or absence of outcome did not affect estimated S.E. Estimated coefficients and intercept from complete-case analysis were comparable to that of the MICE without using the outcome in the imputation model but S.E.'s were larger.

3.5.4 Comparison of methods to tackle missing data

In general the MICE method is the best approach to impute missing data. However, it has been suggested that, in both binary and time-to-event outcomes, when missing rate is low (about 10%), replacement of missing data by median or mean is a reasonable approximation for the MICE. Some examples are as follows.

With binary outcomes, the ability of mean replacement, MI techniques, and complete case analysis were compared [Ambler G et al., 2007]. The data set of Society of Cardiothoracic Surgeons of Great Britain and Ireland was used. The original data included 20378 cases of which 1404 patients had died. Nine variables, out of 14 variables studied involved less than 15% missing values. The missing rate for 3 variables was higher than 20%. Actual missing rates were 20.2%, 42.7%, and 53.4%. In total, 32% had complete data on all variables studied.

Authors found that, under MAR and MCAR mechanisms, performance of the MICE and mean substitution were comparable and much better than complete-case analysis in terms of the proportion of patients classified into the correct risk group and the estimated spearman rank correlation between true and fitted probabilities. Estimated root mean square error (which quantified difference between fitted and true probabilities) for the mean substitution method was marginally higher than that of

the MICE but better than complete-case analysis. Comparing estimated regression coefficients, it has been shown that the MICE produced the lowest level of bias [Ambler G et al., 2007].

In another study, performance of the mean replacement and the MICE, in terms of magnitude of estimated coefficients and Standard Error (S.E.), and discrimination ability were compared [Van Der Heijden GJ et al., 2006].

Data for 398 cases with suspected pulmonary embolism were available of which 246 participants (62%) had complete information on all 26 variables studied. The rates of missing values were as follows: 0% for12 variables, <10% for 11 variables 14% for 1 variable and 21% for 2 variables. The number of variables which contributed to models was 9 for the MICE and 10 for mean replacement. Corresponding figures for discrimination ability was 78.7% and 77.5% respectively. Furthermore, replacement of missing data by mean yielded smaller S.E.'s for 3 variables whereas for 4 variables estimated S.E.'s were the same.

Asia Pacific Cohort Studies Collaborators (APCSC) collects data to identify Coronary Heart Disease (CHD) risk factors. The ability of mean replacement and MI techniques to handle the missing data on a single variable (cholesterol) were compared in 22 studies [Barzi F and Woodward M, 2004]. The cholesterol rate of missing date varied from 0% to 9.1%. Both methods gave similar results in terms of the mean and standard deviation of cholesterol and the estimated coronary mortality hazard ratio. This indicated that when missing rate is low, then application of simpler

methods such as mean replacement might be a very good approximation for complex and sophisticated imputation methods such as the MICE.

3.5.5 Imputation of MNAR data

It has been stressed that 'If missing data are MNAR, valuable information is lost from the data and, there is no universal method of handling the missing data properly' [Donders AR et al., 2006]. It has been noted that by including enough variables in the imputation model the MAR assumption would be more plausible [Van Buuren S et al., 1999]. Furthermore, efficient estimation with MNAR to a great extent depends on prior knowledge about the missing data mechanism [Harel O and Zhou XH, 2007].

When missing data are MNAR, the reason for missingness should be understood and considered into the process of imputation of missing data. As an example, consider a situation in which due to technical problems it is not simple to measure histoscore values below a threshold value such as 5. In this case, it is clear that cases with missing data had a value varies from 0 to 5. Therefore, to impute plausible values, the observed data and the missingness mechanism should be modelled simultaneously.

Another ad hoc approach is to categorise the variable and consider cases with missing data as a single category. This method is simple to implement but lead to loss in information.

3.5.6 Summary

My literature review showed that the MICE method is the best method to impute missing data. However, when missing rate is low application of other methods might produce results comparable to the MICE. Features of this method are compared with other easier approaches (Table 3.7).

Table 3.7: Advantages and disadvantages of methods to tackle missing data

Feature	Complete-case	Median substitution	EM	MICE
No special software is needed	Yes	Yes	Yes	No
Easy to communicate with clinical audience	Yes	Yes	Yes	No
Do not require distributional assumption	Yes	Yes	No	Yes
Preserve data characteristics	No	No	Yes	Yes
Convergence of imputation model is not an issue	Yes	Yes	No	No
Takes imputation uncertainty into account	No	No	No	Yes
Any particular problem	Diminishes the power Gives biased estimated if not MCAR	Artificially reduces the variance	Might give out of range estimates	Requires aggregation of estimates

3.6 Assessing the internal validity of models

3.6.1 Background

Ultimately, the most important issue for a model is its external validity, the extent to which it provides good predictions for similar patients who were not involved in the development of the model. However, before external validity can be checked, it is a prerequisite that there is adequate internal validity. Internal validation refers to the performance in patients from a similar population to those comprising the sample on which the model was developed. Therefore, internal validity is in contrast to external validity, where different populations are used to develop and test the model [Justice AC et al., 1999].

If performance is assessed on the same sample as used for model development, then performance will be overestimated. Internal validation provides an accurate estimate of the upper limit of performance that might be expected for other populations (external validity).

Internal validity can be investigated by splitting the data into training and test samples, doing cross-validation, using Akaike's Information Criterion (AIC), or performing a bootstrap resampling procedure [Harrell FE et al., 1985; Harrell FE et al., 1996].

Data-splitting and cross-validation methods are not appropriate when sample size is low [Steyerberg EW et al., 2001]. Akaike's Information Criterion (AIC) is another

method to tackle overfitting [Collett D, 2003]. A model with higher number of variables provides a better fit than a model with small number of variables. AIC is a trade off between goodness of fit and model complexity. This statistics is defined as below. The lower the AIC, the better the fit is:

AIC= -2log (maximum likelihood) + 3 x (number of parameters)

AIC is asymptotically equivalent to cross-validation but is much faster to implement. Although use of AIC is straightforward, this approach does not test the stability of the form of association. Since in the cohort I am going to analyse EPV is low and one of the aims is to explore possible non-linear effects, I will focus this review on the bootstrap procedure.

3.6.2 Bootstrap procedure

Stepwise variable selection procedures are frequently used in the literature but are not stable as the inclusion or exclusion of a few cases can affect the variables selected for the model and resulting parameter estimates [Sauerbrei W and Schumacher M, 1992; Austin PC and Tu JV, 2004a; Derksen S and Keselman J, 1992].

This issue has been addressed in the development of a prediction model for acute myocardial infraction mortality. In 1000 bootstrap samples, B.E. produced 940 unique models [Austin PC and Tu JV, 2004a]. Out of 29 variables, only 3 variables were significant in all the bootstrap samples, 18 variables were selected in fewer than half of the bootstrap samples, and 6 variables in less than 10%. This demonstrates the sensitivity of B.E. to small differences between bootstrap samples. It has therefore

been recommended to use B.E. in conjunction with bootstrap procedure [Altman DG and Andersen PK, 1989; Harrell FE et al., 1996; Steyerberg EW et al., 2003]. That is, to apply B.E. to a number of bootstrap samples (typically 100) and then to check selection of variables across samples (known as inclusion frequency or percentage).

When modelling across bootstrap samples, the prognostic variables that truly are important should be retained in most models fitted. This is because each bootstrap replication is a random sample that should therefore reflect and mimic the underlying structure of the data, and it is this should drive the variables needed in the majority of models fitted [Altman DG and Andersen PK, 1989; Harrell FE et al., 1996; Steyerberg EW et al., 2003]. Therefore, a measure of inclusion frequency can be used to screen for the selection of the variables [Sauerbrei W and Schumacher M, 1992; Austin PC and Tu JV, 2004b].

The question is what criterion threshold to use for variables to be retained in a model? When the aim is to fit a parsimony model, then only variables with very high inclusion frequency should be retained. On the other hand, when adjustment for covariates is the aim, selection of variables with low inclusion frequency is necessary and therefore, a low value for percentage of inclusion frequency should be selected [Sauerbrei W and Schumacher M, 1992]. The inclusion of a variable in the model at selection levels of 1% and 5% in the original data can be checked against a cut of value for the bootstrap inclusion fraction of 73% and 50% respectively [Sauerbrei W and Schumacher M, 1992].

3.7 Combining methods to develop a prognostic model

3.7.1 Research on combination of methods

In a recent study explored the issue of combining methods in prognostic studies [Heymans MW et al., 2007]. The study population consisted of 628 patients and the clinical aim of the study was to determine prognostic variables for low back pain. The numbers of variables and events were 31 and 135 respectively (EPV=4.4). Missing data rate per variable ranged from 0% (for 2 variables) to 48.1%.

In total 4 aspects were explored, fitting 20000 models. To develop the multifactorial model, a B.E. procedure with the probability to remove of 0.50 was applied to:

- 100 imputed data sets (MI_{1to100})
- 200 bootstrap samples drawn from the first imputed data set (MI_1 x B)
- 20000 samples (200 bootstrap samples from each of 100 imputed data sets) ($MI_{j(j=1,..,100)}$ x B)
- 2000 samples (200 bootstrap samples from each of first 10 imputed data sets) ($MI_{j(j=1,..,10)}$ x B)

To impute the data, all 31 variables were offered to the imputation program, but the underlying model failed to converge due to multicollinearity and computational problems. Therefore, imputation was undertaken separately for each variable. To

impute missing data for a variable, only other 'complete-data' variables and variables which had correlation higher than 20% with the variable to be imputed were included in the imputation run. Thus, a series of imputation runs were undertaken, consisting of 10 to 25 variables per run. No further information was given about number of variables used for each variable to impute missing data. Although it has been advised that imputation of 10 data sets is enough [Schafer JL, 1999], missing data were imputed 100 times to be able to estimate inclusion frequency per variable precisely.

The number of times that any of the 31 variables appeared in the multifactorial model was recorded. For each of the four scenarios, the number of variables selected in at least 60% of models was reported (Table 3.8).

As summarised in Table 3.8, for MI_{1to100} inclusion frequency varied from 27% to 100% whereas for the MI_1 x B (200 bootstrap samples from first imputed data set) the range was 51.8% to 100%. Authors concluded that MI_{1to100} was more specific in distinguishing variables with very high and very low inclusion frequency [Heymans MW et al., 2007].

However, this conclusion might not be right in all situations. As explained above the first imputed data set was selected to draw 200 bootstrap samples. Authors did not reveal the selection of variables in that particular imputed data set. In the case where all variables were important in that data case, it might not be surprising for the inclusion frequency to be large across the bootstrap samples. In my opinion, since the rate of missing values was high, imputed data sets were more likely to vary than bootstrap samples drawn from one imputed data set. Results might be different if

another imputed data set had been selected at the start. Furthermore, in development of the model, a large P-value of 0.50 was used which might result in numerous covariates and unstable models. Results were not compared with of a more conventional P-value of 0.05.

Authors noted that combined models, which consider both imputation and sampling variations, ($MI_{j(j=1,..,100)}$ x B and $MI_{j(j=1,..,10)}$ x B) were similar to MI_{1to100} in terms of order of selection of variables (data not shown) but were similar to MI_1 x B in terms of range of inclusion frequency. Therefore, it was concluded that missing data variation had a larger impact on inclusion frequency than sampling variation.

This conclusion might also be in doubt. As explained above, comparison of combined models with MI_1 x B might lead to a different conclusion if another imputed data set had been selected. Comparisons with MI_1 x B might be very sensitive to the imputed data set selected.

Finally, a combination of bootstrap with 10 imputed data sets worked as well as with 100 imputed data sets. They have found that using 10 and 100 imputed data sets resulted in similar selection of the variables and therefore imputation of 10 data sets should be adequate [Heymans MW et al., 2007].

As a sensitivity analysis, only $MI_{j(j=1,..,10)}$ x B was repeated with a moderate P-value of 0.157. This P-value (0.157) was chosen since Sauerbrei et al. argued that Backward Elimination (B.E.) variable selection approach at a 0.157 level is a good

approximation for all subset regression with Akaike's Information Criterion (AIC) [Sauerbrei W, 1999].

Applying this P-value (0.157), results were markedly different. The range of selection of variables varied from 19.2% to 99.1% indicating a very high sensitivity of aspects in relation to a nominal P-value set (Table 3.8). With a P-value of 0.157, only 4 variables were retained in more than 60% of samples. The range of selection of variables was similar to MI_{1to100} where sampling variation was not taken into account.

Table 3.8: Range of selection of variables and number of variables retained in more than 60% of samples in Heymans *et al.* study (2007)

Model	Number of samples	P-value to remove	Aspect investigated	Range of selection of 31 variables across models	Number of variables selected in > 60% of samples
MI_{1to100}	100	0.50	Imputation variation with high P-value	27.0%, 100%	13
MI_1 x B	200	0.50	Sampling variation with high P-value	51.8%, 100%	18
$MI_{j(j=1,..,100)}$ x B	20000	0.50	Both high P-value and large number of imputations	57.1%, 99.4%	27
$MI_{j(j=1,..,10)}$ x B	2000	0.50	Both high P-value and small number of imputations	55.1%, 99.5%	26
$MI_{j(j=1,..,10)}$ x B	2000	0.157	Both moderate P-value and small number of imputations	19.2%, 99.1%	4

3.7.2 Overview

Heymans *et al.* fitted 20000 B.E. logistic regression models to explore 4 different aspects: only taking into account imputation uncertainty, only taking into account imputation variation, or allowing for both [Heymans MW et al., 2007]. Although their experiment had several limitations, it has the advantage that it allows for two sources of variation: variation due to imputation of missing data and sampling variation.

However, in my opinion, the most important limitations of the study was that authors assumed that effect of continuous variables was linear. Attempt was not made to check whether data comply with this important assumption, nor whether modelling would be better if a different form was used. Application of traditional linear regression models might result in loss of useful predictors which has non-linear association with the outcome.

Chapter 4 DESIGN AND METHODS

4.1 Aims and objectives

The key problems in prognostic modelling are to restrict number of variables being offered to the multifactorial model, to assess the optimum form of association, to avoid attrition in the sample size, and to assess the internal validity of the model.

The main aim of this project is to develop statistical methods for modelling an outcome in a survival analysis when there are many potential variables. This requires a reduction in the number of variables, sufficient to allow model fitting that involves all potentially informative variables, while guarding against overfitting and instability in the model.

The main statistical objectives are given below, together with an indication of where in the thesis corresponding results are reported.

i) Preliminary work prior to the main modelling

1. Describe the distribution of variables in the data set (Chapter 5)
2. Calculate standard Nottingham Prognostic Index (NPI) for the patients, categorise patients into risk groups by applying a range of cut offs to these index scores, and refit NPI using the data set for this research (Chapter 6)

ii) Main modelling

1. Compare methods for detecting form of association (Chapter 7)
2. Develop pragmatic strategies for fitting of multifactorial models for data sets comprising many skewed variables and missing values (Chapter 8)

iii) Additional analyses to enhance understanding

Explore details of methods applied by means of further investigations of elements of the procedure:

- Imputation of missing data in Chapter 9
- Selection of form of association in Chapter 10

In addition, the clinical objectives of this thesis are: to identify biomarkers with potential to inform prognosis, to develop a prognostic model for risk stratification of breast cancer patients, and to detect a low-risk group for clinicians to aid a treatment strategy that avoids harsh treatments.

4.2 Overall study design

The research is based on a secondary analysis of an existing cohort of Estrogen Receptor positive (ER+) tamoxifen treated breast cancer patients.

4.3 Data set

John Bartlett is a professor of biology with special interest in molecular profiling of tamoxifen resistance in breast cancer. While he was working in Glasgow University, he and his colleagues began to develop a data set to study the pathway of tumour progression. They utilised Tissue Microarray Analysis (TMA) of tumour tissues to collect data on a large number of biomarkers, so as to identify biomarkers which can provide additional risk factors for breast cancer. His collaborators were Tove Kirkegaard, Liane McGlynn, Sian Tovey, and Timothy Cooke.

TMA has been taken up by research institutions around the world, in particular those involved in cancer research [Chen W and Foran DJ, 2006]. In TMA, tissue is embedded in paraffin. Then using a hollow needle tissue cores are removed from the region of interest, a process similar to clinical biopsies. The diameter of tumours removed is as small as 6 millimetres. Tissue cores are then inserted into a paraffin block. Microtome, which is a mechanical instrument, is used to cut sections from blocks for microscope examination. Staining is frequently used in biology to enhance contrast in the microscopic image. Staining is scored according to the cellular distribution of expression of individual markers and each core is analysed separately for membrane, cytoplasmic and nuclear localisation of biomarkers. Incorporating the

intensity and percentage of positive staining, histoscore values are calculated which usually range from 0 to 300 and units are on the whole arbitrary.

i) Patients and number of events

The sample was comprised 401 ER+ women diagnosed during 1983 and 1999 at Glasgow Royal Infirmary formed the sample. Median follow-up time was 6.16 years and all patients received tamoxifen for some of the follow-up time (median of 5 years).

At the end of follow-up there had been 112 recurrences, and in 84 of these cases the patient was still was on tamoxifen treatment at the time of recurrence. A total of 74 of the patients with a recurrence died during the follow up period.

ii) Outcomes studied

The primary outcome for the study is Recurrence Free Survival (RFS) and secondary outcomes are Recurrence Free on Tamoxifen treatment (RFoT) and Overall Survival (OS). Prognostic models are developed only to predict RFS, but Kaplan-Meier (K-M) curves are presented for all end points in relation to risk groups obtained from the models developed for RFS.

iii) Variables

Data for 72 tissue microarray variables describing 41 protein biomarkers were available. In addition, there are three clinical variables (nodal status with 3 levels, Grade with 3 levels, and pathological tumour size) and age at diagnosis. The clinical variables were measured as follows. Tumour size was based on measurement of the

mastectomy specimen. Histological grade (1 to 3) was determined based on criteria of Bloom and Richardson [Bloom HJ and Richardson WW, 1957]. The Bloom-Richardson grading method is based on three features of invasive breast cancers: the percentage cancer composed of tubular structures, the rate of cell division, and the nuclear pleomorphism of tumor cells (nuclear grade, change in cell size and uniformity). Each of these 3 features is rated from 1 to 3. Summation of these scores, which give a total score that ranges from 3 to 9, is used to grade the tumours as follows:

- Grade 1 tumor (well-differentiated): scores 3 to 5
- Grade 2 tumor (moderately-differentiated): scores 6 to 7
- Grade 3 tumor (poorly-differentiated): scores 8 to 9

Lymph node involvement was determined based on biopsy of a lower axillary node, an apical axillary node, and a node from the internal mammary chain. Patients were staged into 3 groups in terms of lymph node findings:

- Stage 1: Tumour absent from all 3 nodes sampled
- Stage 2: Tumour in low axillary node only.
- Stage 3: Tumour in either of apical or internal mammary nodes

4.4 Overview of general methods of analyses

General methods which are relevant to the whole thesis are presented here and an outline of methods used in one or another chapter. More detail of methods specific to a single results chapter is given in that chapter. These are:

- Ascertain form of association (Chapter 7)
- Bootstrap sampling and investigation of stability of forms (Chapter 8)
- Median replacement of missing data (Chapter 9)
- Use of dichotomised biomarkers in the multifactorial model (Chapter 10)

Methods relevant to more than one chapter, and described here are:

- Cox regression model and check for Proportionality of Hazard assumption (4.4.1)
- Fractional Polynomial modelling (4.4.2)
- Minimum P-value method (4.4.3)
- Multiple imputation of missing data (4.4.4), including:
    - Aggregate model results across imputed data sets and calculate a single risk score
    - Apply Backward Elimination (B.E.) variable selection method in the case of multiply imputed data sets
    - Calculate Hazard Ratios and Confidence Intervals (C.I.) in the case of multiply imputed data sets
- Stratification of patients into risk groups by means of risk scores (4.4.5)
- Graphical display of survival for risk groups (4.4.6)
- Comparison of performance of models (4.4.7)

4.4.1 Cox regression model and Proportional Hazard (PH) assumption

The application of the linear Cox model is considered the basic and standard tool in the field of time-to-event data analysis [Cox DR, 1972]. The model requires Proportionality of Hazards (PH) which means that the survival advantage is constant across time, from the first months through to the last years of follow-up [Sposto R, 2002]. To check whether data holds with the PH assumption, I will plot the Schoenfeld residual versus survival time [Hess KR, 1995].

Another option to deal with time-to-event data is to apply parametric survival models such as Weibull or log-logistic regression model [Orbe J et al., 2002; Carroll KJ, 2003].Parametric survival models allow for a wider set of inferences to be made and provide insight into the shape of the baseline hazard [Nardi A and Schemper M, 2003]. However, evidence for the merit of a chosen parametric model is needed [Therneau TM and Grambsch PM, 2000]. Given the relative lack of interest in these models among the clinical collaborators I have decided not to apply parametric survival models.

4.4.2 Fractional Polynomial Modelling

Fractional Polynomial (FP) modelling is a univariate technique that explores the data to find optimum power transformation for a variable [Royston P and Sauerbrei W, 2008]. There are two classes of FP: first degree (FP1), and second degree (FP2) fractional polynomials.

The first degree Fractional Polynomial technique (FP1), performing 8 tests, checks whether fit is improved by a power transformation of the variable X, X^p, where p is chosen from S= {-2,-1,-0.5, 0, 0.5, 1, 2, 3}. FP with value of p=1 is synonymous with a linear regression and p=0 indicates that a logarithmic transformation is needed for optimum linear modelling of a risk factor.

A polynomial model of degree 2 (FP2) is an extension to $\beta_1 X^{p_1} + \beta_2 X^{p_2}$ which compares 36 different power combinations. It can be seen that ($p_1 = 1, p_2 = 2$) is equivalent to quadratic regression. The case $p_1 = p_2$ is known as a repeated power model and has been defined as $\beta_1 X^p + \beta_2 X^p LnX$ [Royston P and Altman DG, 1994]. The power selection procedure has been chosen empirically as that which maintains approximately the correct type one error, as has been shown in a simulation study [Ambler G and Royston P, 2001]. For each variable separately, the following steps are carried out:

i) Is the variable is needed at all?

Fit the best FP2 model and test it versus null model using 4 degrees of freedom (d.f.). If it is not significant drop the variable and stop. If it is significant go to the next step.

ii) Simplification

- Test best FP2 versus a linear fit using 3 d.f. If it is not significant declare the final model to be a straight line and stop. If it is significant go to next step.
- Test best FP2 versus best FP1 (p≠1) using 2 d.f. If the test is significant declare the final model to be FP2. Otherwise, the best model would be the best FP1.

Multifactorial Fractional Polynomial (MFP) modelling is an extension of FP to check whether power transformation is required in the multifactorial model. MFP, after fitting of linear factors, ascertains whether model fit would be improved by using a polynomial form for any of the linear variables.

4.4.3 Minimum P-value method

Minimum P-value method is a univariate technique which is used to detect threshold effects. This method explores the data to find the optimal split for a continuous variable. After a systematic search across all possible values, the value chosen as the cut point will be that with the smallest corresponding P-value in a Log-Rank test, when comparing the survival curve of the two groups formed [Lausen B and Schumacher M, 1992].

To avoid groups with small numbers of patients, no split at the outer 20% of distribution of variables will be applied (lowest and highest 10%). Furthermore, the Altman formula will be applied to correct for multiple comparisons undertaken [Altman DG et al., 1994]. The adjusted P-value will be calculated applying formula below:

$$P_{alt10} = -1.63 P_{\min}(1 + 2.35 Ln(P_{\min}))$$

4.4.4 Imputation of missing data

My literature review showed that Multivariable Imputations by Chained Equations (MICE) method is the best approach to impute missing data. Furthermore, median substitution is a good approximation for the MICE method when missing rate is about 10%. Although the rate of missing values in the current data set is not high, I will use the MICE method to impute 10 data sets [Schafer JL, 1999].

Specification of the imputation model is the first steps. Predictive Mean Matching (PMM), polytomous regression, and logistic regression will be used to impute missing data for continuous, categorical, and binary data respectively. In the PMM method, the complete-case whose value is closest to the imputed value is chosen. It takes the observation from the complete-case as the imputed value.

The second step is to select a set of variables to enter into the imputation model. Using all available information make the MAR assumption more plausible. However, in presence of a dozen of covariates, it is not feasible nor necessary, due to multicollinearity and computational problems [Van Buuren S et al., 1999]. I will use a reduced set of informative variables or family sets of biomarkers in the imputation model (see Chapter 8). This will then be challenged by including a large number of variables in the imputation model (see Chapter 9). Furthermore, outcome is included [Moons KG et al., 2006].

The third step is to draw the imputation values. To impute the missing value of X_j, a regression model relates X_j to other variables in the imputation model. This

regression model is then used to create imputed values by drawing from the posterior predictive distribution. Each predictor with missing values is considered in turn using the current imputed values for each of the other predictors [Van Buuren S et al., 1999]. This iteration process ends when all variables have been updated [Clark TG and Altman DG, 2003]. If $X = (X_1, X_2, ..., X_k)$ are k random variables where each variable contains missing data and t represents the iteration number, missing data are imputed from the following sequence of Gibbs sampler iterations, as explained below [Van Buuren S and Oudshoorn K, 2000]:

For X_1 draw imputations X_1^{t+1} from $P(X_1 \left| X_2^t, X_3^t, ..., X_k^t \right.)$
For X_2 draw imputations X_2^{t+1} from $P(X_2 \left| X_1^{t+1}, X_3^t, ..., X_k^t \right.)$
…..
For X_k draw imputations X_k^{t+1} from $P(X_k \left| X_1^{t+1}, X_2^{t+1}, ..., X_{k-1}^{t+1} \right.)$

This entire process is repeated and the imputed values which are created at the 5th round will be used as the first imputed data set. The whole process explained will be repeated 10 times to replace each missing data by 10 values, thus creating 10 data sets [Van Buuren S et al., 1999].

i) Aggregation of estimates across imputed data sets

The creation of 10 data sets means there is a requirement for 10 modelling analyses, one for each data set, and there will therefore be 10 different estimates for each parameter. Estimates derived from imputed data sets therefore need to be combined and this will be achieved applying Rubin's rule [Rubin DB, 1978]. The coefficients and standard errors will be aggregated across the imputed data sets by Rubin's

formulae [Rubin DB, 1978], where $\hat{\beta}_i$ is the estimated regression coefficient and $\hat{\beta}$ is the aggregated coefficient (M=10 for my application).

$$\hat{\beta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_i \qquad \hat{Var}(\hat{\beta}) = \frac{1}{M}\sum_{i=1}^{M}\hat{Var}(\hat{\beta}_i) + (1+\frac{1}{M})(\frac{1}{M-1})\sum_{i=1}^{M}(\hat{\beta}_i - \hat{\beta})^2$$

ii) Backward Elimination (B.E.) variable selection method with multiply imputed data sets

If a single multifactorial model is being developed then application of B.E. is straightforward. However, when there are 10 imputed data sets, B.E. will not directly be feasible. The 10 models developed are likely to retain different variables and aggregation of estimates (as per 4.4.4 (i)) will not be possible.

To simplify the full model, by B.E. of imputed data sets, a series of iterative steps is required. At each step, the results are aggregated across the 10 data sets, and the variable with the highest P-value (exceeding 0.05) is removed. Another set of 10 models is fitted with remaining variables, results are aggregated, and P-value assessed for a variable to drop (if P-value >0.05). The whole process continues until all variables remain significant in more than 5 data sets [Van Buuren S et al., 1999; Clark TG and Altman DG, 2003].

iii) Calculation of Hazard Ratios (HR) and Confidence Intervals (C.I.)

Hazard Ratios (HR) and corresponding 95% Confidence Intervals (C.I.) will be calculated from regression coefficients and standard errors that have been imputed across multiply imputed data sets.

iv) Calculation of a risk score

A risk score will be calculated for each of 10 imputed data sets. For each patient a single averaged risk score will be calculated by averaging her estimated risk scores from each of the 10 imputed data sets.

4.4.5 Formation of risk groups

When wishing to categorise a prognostic risk score into risk groups, the important issues are to create reasonable number of risk groups, each containing an adequate number of patients from each risk group, and to select appropriate cut offs.

In development of NPI, cut points were chosen that created 3 risk groups and as summarised in Table 2.2, this generally creates risk groups containing unequal numbers of patients. However, for the biomarker models I was to develop, I decided to create 4 risk groups from risk scores obtained. This is because one of the clinical ambitions of this study was to identify a subset of very low risk patients. Such a low risk group is unlikely to be identified unless the group is a relatively small proportion of the population of breast cancer patients.

Selection of a very low cut off might guarantee detection of a risk group with sufficiently low risk of recurrence. However, the cost is that the estimated rates might not be robust due to low sample size in the risk group. Therefore, I will selet the cut off so as to create a low risk group containing 25% of the whole data.

One way to define 4 groups would be to plot the distribution of risk score and select cut offs which guarantee the most diverged risk groups. However, this data-

dependent approach reduced the generalisability of estimates obtained to other data sets. Instead, I will created a categorised 4-level variable by using as cut offs the three quartiles of risk score. This makes the place of split blind to the distribution of risk score. Furthermore, each group will contain roughly 100 patients (25% of data) which is a reasonable figure.

4.4.6 Graphical display of risk groups and calculation of survival rates

For graphical display of performance of risk groups derived in risk stratification, Kaplan Meier (K-M) curves corresponding to risk groups created will be plotted. Number of patients at start and followed for 3, 5, 7, 9, and 10 years are reported below each graph.

Furthermore, 5, 7, and 10-year RFS rates in the lowest and highest risk groups are reported. Details for calculation of 7-year RFS are given. Corresponding figures for 5 and 10 years are calculated in a similar fashion.

Based on the definition the survival function, S (7), is the probability of being alive at least till the 7^{th} year of follow up. Therefore, survival at the 7^{th} year depends on survival at first, second…and 6^{th} year which implies that $S(7) = P(T \geq 7)$. In the actuarial life-table procedure, the whole follow-up duration will be split into intervals (an example of 1 year intervals is (0, 1], (1, 2], (2, 3]… and (6, 7]). If n_i and d_i show number of patients at risk just before the i-th interval and the number of events

at i-th interval, then the probability of surviving to 7th year is given by

$$S(7) = \prod_{i=1}^{7} (1 - \frac{d_i}{n_i})$$

4.4.7 Comparison of models

The ongoing discovery of new risk factors poses new questions on how best to quantify the contribution of added risk factor(s) to improving risk prediction [Pencina MJ et al., 2008]. However, Altman *et al.* noted that the appropriate definition of prognostic performance is open to debate [Altman DG and Royston P, 2000].

The Area Under Curve (C-index), Nagelkerke R-square (predictive ability), and Likelihood Ratio Test (LRT) are the most frequently used statistics in the literature, for comparing the performance of different models or to quantify the improvement in model accuracy when adding a new risk factor to set of standard predictors [Hanley JA and McNeil BJ, 1982; Pencina MJ and D'Agostino RB, 2004]. In this thesis I will use these 3 statistically-oriented statistics.

However, one important limitation with C-index is that a very large independent association of a new marker with an outcome is required to result in an incrementally larger AUC [Pepe MS et al., 2004; Greenland P and O'Malley PG, 2005; Ware JH, 2006]. It has been shown that an OR as large as 3 may have little impact on C-index [Pepe MS et al., 2004]. Furthermore, these statistics are not necessarily informative in clinical practice. Therefore, I also will report clinically-oriented statistics: Net

Reclassification Index (NR Index), estimated RFS rates in the lowest and highest risk groups, and Prognostic SEParation (PSEP).

i) Statistically oriented statistics

Predictive ability (R-square)

Predictive ability is addressed by Nagelkerke R-square. This statistic, which varies between (0, 1), describes the ability of a model to an predict outcome. Values of 0 and 1 indicates very poor and very high predictive ability respectively [Harrell FE, 2001].

Discrimination ability (C-index)

Discrimination refers to the ability to separate patients with different responses [Justice AC et al., 1999]. Discrimination is measured using Harrell's C-index (concordance index) which is a generalisation of Area Under Curve (AUC). This statistic varies between 0.5 and 1 where values near 1 indicate high discriminatory power.

Goodness of fit (Likelihood Ratio Test)

For all models, I will report Likelihood Ratio Test (LRT) which indicates how well the model fits the data.

Calibration

Calibration quantifies the extent to which predicted probabilities match observed probabilities. Such statistics show the ability of the model to make unbiased

estimates of outcome [Clark TG and Altman DG, 2003]. In addition there are other statistics that can be used to compare the performance of models such as D statistics [Royston P and Sauerbrei W, 2004] and Brier score which quantifies the mean square error of prediction [Graf E et al., 1999].

I will not calculate the calibration of models. This is because one of my particular aims is to detect a subgroup of patients with sufficiently low risk of recurrence and therefore the calibration of the model over the entire range of the prognostic risk is not directly relevant. Furthermore, Harrell *et al.* noted that discrimination of a model is of primary concern [Harrell FE et al., 1996]. Per discussions with clinical collaborators of this study, I will only report C-index, R-square, and LRT.

ii) Clinically oriented statistics

Net Reclassification Index (NR Index)

A recent paper suggested the use of Net Reclassification Index (NR Index) as a tool to assess usefulness of a new risk factor [Pencina MJ et al., 2008]. NR Index is much more sensitive to the addition of a new risk factor than other statistics described.

This statistic checks the extent to which the addition of new variables to a standard model reclassifies patients into more appropriate risk groups. This method considers the joint distribution of patients into risk groups, by the two risk grouping schemes being compared. For patients who did and did not experience the event, it quantifies the 'correct' movement in the risk group classifications i.e. upwards for patients who did experience the event and downwards for patients who did not [Pencina MJ et al., 2008].

Net gain in cases *with event* has been defined as the difference in proportion of subjects who moved into a higher or lower risk group. The reverse calculations will be made for *event free cases*. NR Index is defined as summation of net gains [Pencina MJ et al., 2008]. Statistical significance of net gains can be checked following logic for McNemar's test in correlated proportions [Pencina MJ et al., 2008]. Simple tests are developed to ascertain the significance of net gain in recurred and non-recurred cases and also NR Index. To calculate the NR Index, risk groups derived from NPI by applying quartiles as cut offs will be used as the risk grouping scheme.

RFS rate in the lowest and highest risk groups

Model performance will be also assessed in terms of RFS rate in lowest risk group, calculated as explained in section 4.4.6. Prior to the start of the research, I discussed this issue with the clinical collaborators of this study. A minimum 10-year RFS of 95% was proposed. Detection of the high risk patients is also clinically important because they are deemed to need aggressive therapy. Therefore, RFS rates for highest risk groups will also be reported.

Since the number of patients in the cohort who have been followed for 10 years or more is very low (about 11%), an estimate at this time point might not be robust. Therefore, to compare the ability of model to identify low risk patients, actuarial 7-year RFS (with 40% follow-up data) will be used.

Prognostic SEParation (PSEP)

Prognostic SEParation is a general concept, with the precise definition depends on context [Altman DG and Royston P, 2000]. For the purpose of this research, I will use PSEP to describe the separation between low and high risk groups. I specify Prognostic SEParation (PSEP) as the difference at 7-year RFS of lowest and highest risk groups. I also will report corresponding figures at 5 and 10 years.

4.5 Software

Bar charts will be plotted using Microsoft EXCEL software. K-M curves and life-table analysis (to estimate actuarial event free rates) are produced using Statistical Package for Social Science (SPSS version 14).

A series of packages which work under R software (version 2.5.1) will be used [R Development Core Team, 2007]. To detect polynomial effects and to develop Multivariate Fractional Polynomial models, MFP package will be used [Ambler G and Benner, 2008]. Maxstat routines will be used to run minimum P-value method process so as to detect threshold effects [Horton T, 2007]. Missing data will be imputed using the MICE package [Van Buuren S and Oudshoorn C.G.M., 2007]. Estimated regression coefficients and standard errors will be combined across imputed data sets using Mitools library [Lumley.T., 2008]. Tree-based Survival Models will be applied using rpart package [Therneau TM and Atkinson B, 2009]. Performance of models (discrimination and predictive ability) will be assessed using Design [Harrell FE, 2008] library. NR Index will be calculated manually.

Chapter 5 DESCRIPTION OF BIOMARKERS AND CLINICAL VARIABLES

5.1 Introduction

The Glasgow data set is a typical example of a data set with a large number of biomarkers. In this Chapter descriptive statistics for all of the 72 biological variables is reported.

5.2 Methods

For each biomarker, descriptive statistics (minimum (Min), maximum (Max), and quartiles (Q1, Q2, and Q3)) plus number and percentage of missing data (Missing)

are reported. For additional insight to distribution, histograms of some biomarkers are plotted. Biomarkers with skewed distribution are flagged with * sign.

Professor John Bartlett, using biological expertise on the basis of presumed role in the pathway to cancer progression, divided biomarkers into seven substantive biomarker family sets; AKT, BAD, PgR, RAS, MTOR, MAPK, HER. Biomarkers not included in these families formed an eighth group, named 'Non-family' set. Statistics for each biomarker set and also for clinical predictors are summarised separately. For each biomarker, an abbreviation is selected to be used throughout this thesis.

5.3 Results

A list of all 72 biomarkers and clinical variables, in alphabetic order, and the section in which each summarised is given in Table 5.1.

Table 5.1: List of all biomarkers and section statistics are given

Abbreviation in thesis	Family	Section	Abbreviation in thesis	Family	Section
Aib1	HER	5.3.7	P118cy	PgR	5.3.6
Aibfis1	HER	5.3.7	P118me	PgR	5.3.6
Aibfis2	HER	5.3.7	P118nu	PgR	5.3.6
Akt1cy	AKT	5.3.1	P167cy	PgR	5.3.6
Akt1nu	AKT	5.3.1	P167me	PgR	5.3.6
Akt2cy	AKT	5.3.1	P167nu	PgR	5.3.6
Akt3cy	AKT	5.3.1	Pakt1cy	AKT	5.3.1
Badcy	BAD	5.3.2	Pakt1nu	AKT	5.3.1
Baxcy	BAD	5.3.2	Pakt2cy	AKT	5.3.1
Bcl2	BAD	5.3.2	Pakt2nu	AKT	5.3.1
Bclxl	BAD	5.3.2	Panaktcy	AKT	5.3.1
Egfrmax	HER	5.3.7	Panaktnu	AKT	5.3.1
Erbcy	PgR	5.3.6	Pbad112c	BAD	5.3.2
Erbnu	PgR	5.3.6	Pher2cy	HER	5.3.7
Ercy	PgR	5.3.6	Pher2me	HER	5.3.7
Erhisto	PgR	5.3.6	Pher2nu	HER	5.3.7
H4hfr1cy	HER	5.3.7	Pmapkcy	MAPK	5.3.5
H4hfr1me	HER	5.3.7	Pmapknu	MAPK	5.3.5
H4hfr1nu	HER	5.3.7	Pmtor	MTOR	5.3.4
H4jrcy	HER	5.3.7	Pp70s6k3	MTOR	5.3.4
H4jrme	HER	5.3.7	Praf259cyy	MAPK	5.3.5
H4jrnu	HER	5.3.7	Praf259nu	MAPK	5.3.5
Her2fish	HER	5.3.7	Praf338cy	MAPK	5.3.5
Her2me	HER	5.3.7	Praf338nu	MAPK	5.3.5
Hrascy	RAS	5.3.3	Prhisto	PgR	5.3.6
Hrasnu	RAS	5.3.3	Ptency	MTOR	5.3.4
Jrh3cy	HER	5.3.7	Ptennu	MTOR	5.3.4
Jrh3me	HER	5.3.7	Raf1cy	MAPK	5.3.5
Jrh3nu	HER	5.3.7	Raf1nu	MAPK	5.3.5
Krascy	RAS	5.3.3	Rkipcy	Non-family	5.3.8
Krasnu	RAS	5.3.3	Rkipnu	Non-family	5.3.8
Mtor	MTOR	5.3.4	Tace	Non-family	5.3.8
Mapkcy	MAPK	5.3.5	Tacep	Non-family	5.3.8
Mapknu	MAPK	5.3.5	Tescy	Non-family	5.3.8
Nrascy	RAS	5.3.3	Tesnu	Non-family	5.3.8
Nrasnu	RAS	5.3.3	Tunel	Non-family	5.3.8

5.3.1 AKT family

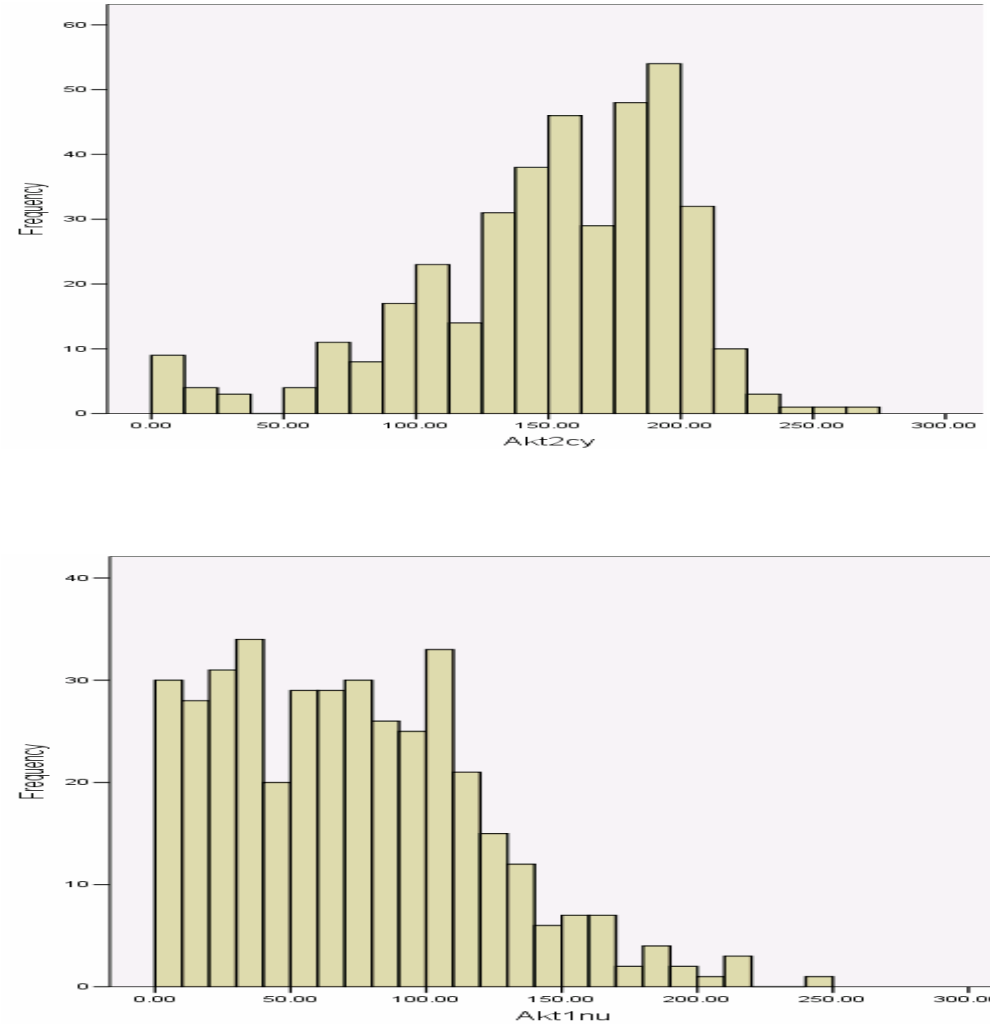
Ten biomarkers forms this family set. Akt1cy and Akt1nu had the lowest missing rate (1 per cent) while data was not available for 4.7% of patients on Panaktcy and Panaktnu (Table 5.2). A total of 356 patients (89%) had data available on all 10 biomarkers.

Table 5.2: Distributional statistics and rate of missing values for AKT family set

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
AKT1	Cytoplasmic	Akt1cy	0	80	112	133	225	1%
	Nuclear	Akt1nu *	0	31	67	102	250	1%
AKT2	Cytoplasmic	Akt2cy	0	125	158	188	275	4%
AKT3	Cytoplasmic	Akt3cy	0	65	92	113	180	5%
Phospho AKT 308	Cytoplasmic	Pakt1cy	0	47	78	100	190	2%
	Nuclear	Pakt1nu *	0	8	20	38	140	2%
Phospho AKT 473	Cytoplasmic	Pakt2cy	0	37	73	110	200	3%
	Nuclear	Pakt2nu *	0	13	27	53	180	3%
Pan AKT	Cytoplasmic	Panaktcy	0	60	85	109	210	5%
	Nuclear	Panaktnu *	0	15	35	64	160	5%

While the distributions of nuclear expression histoscores (4 biomarkers) were highly skewed, the distributions of cytoplasmic histoscores (6 biomarkers) were not far from normal distribution. Histograms of Akt2cy and Akt1nu are given as an example (Figure 5.1).

Figure 5.1: Examples of distribution of biomarkers in AKT family (Akt2cy top panel and Pakt2cy bottom panel)



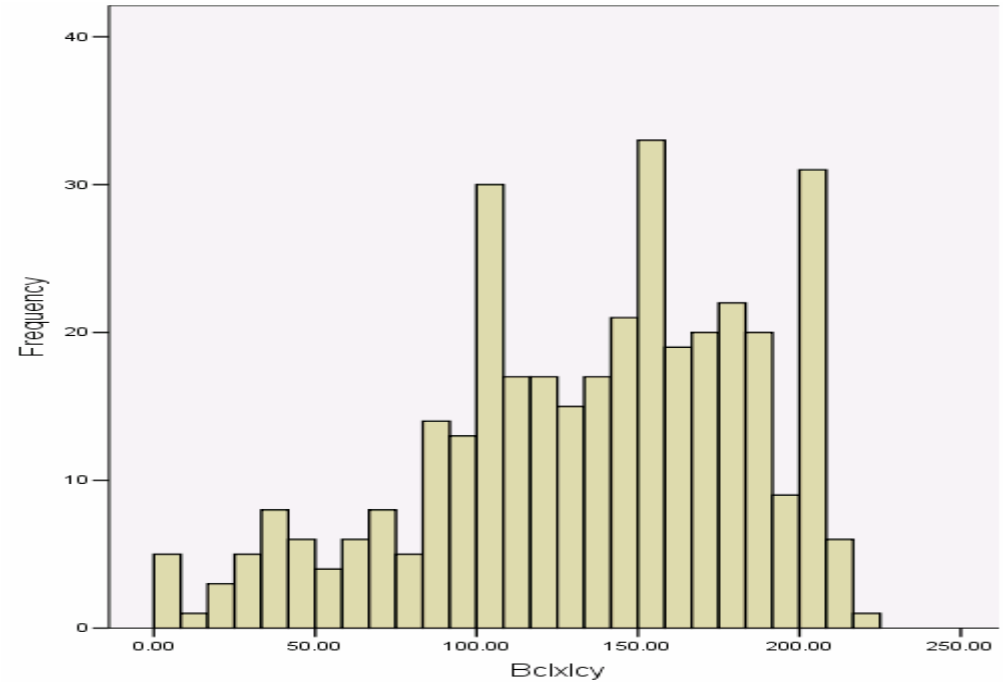
5.3.2 BAD family

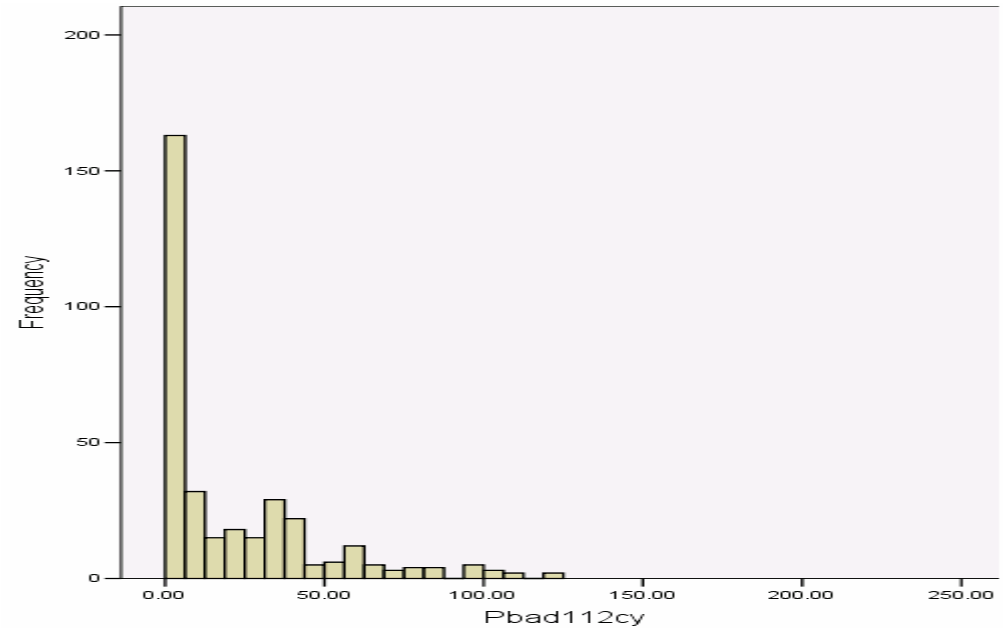
Five biomarkers forms this family set. Only cytoplasmic expression histoscores were available. All of these biomarkers had 7% or higher missing rate, where the highest missing rate was 14% (Table 5.3). A total of 294 patients (73%) had data available on all 5 biomarkers. The distributions of all biomarkers were skewed (see examples of negative and positive skewed distributions in Figure 5.2).

Table 5.3: Distributional statistics and rate of missing values for BAD family set

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
BAD	Cytoplasmic	Badcy *	0	37	67	100	202	12%
BAX	Cytoplasmic	Baxcy *	0	7	27	50	153	13%
Bcl-2	Cytoplasmic	Bcl2cy *	0	20	53	105	200	7%
Bcl-xl	Cytoplasmic	Bclxlcy *	0	104	143	175	220	11%
Phospho BAD 112	Cytoplasmic	Pbad112cy *	0	0	7	34	122	14%

Figure 5.2: Examples of distribution of biomarkers in BAD family (Bclxlcy top panel and Pbad112cy bottom panel)





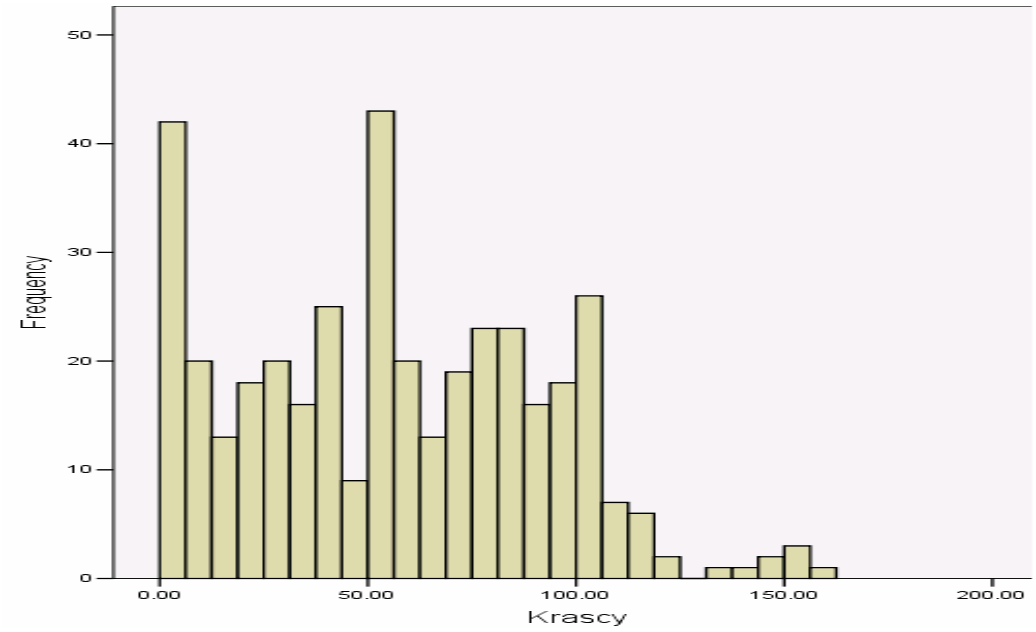
5.3.3 RAS family

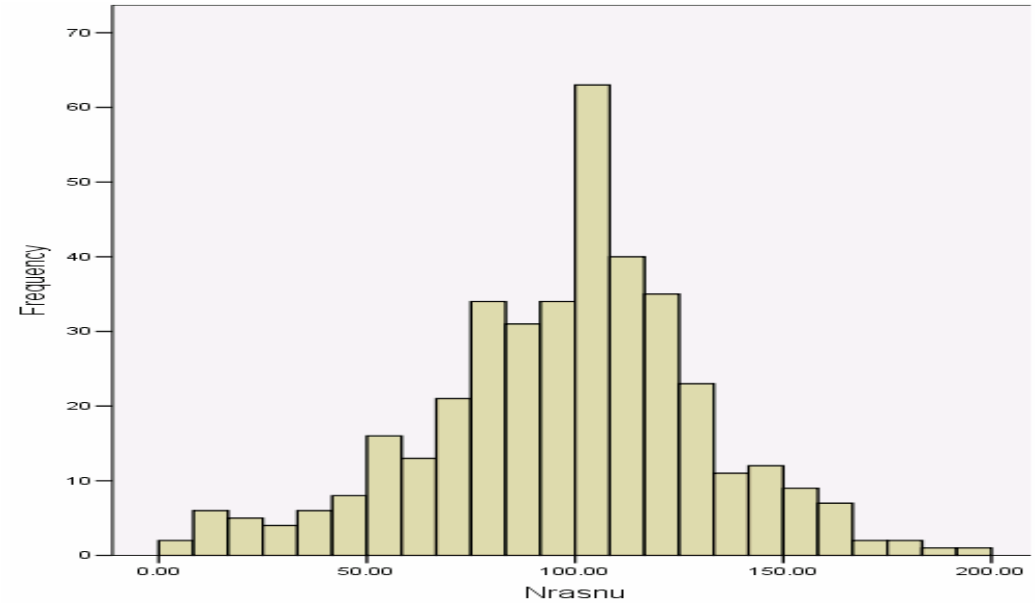
The RAS family comprised 6 biomarkers none of them had a missing rate higher than 4% (Table 5.4). In total, 372 patients (93%) had data available on all 6 biomarkers. The distribution of Hrascy, Nrascy, and Nrasnu was not far from normal. Distributions of other 3 biomarkers were positively skewed. Examples are given in Figure 5.3.

Table 5.4: Distributional statistics and rate of missing values for RAS family set

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
H-Ras	Cytoplasmic	Hrascy	0	90	125	157	232	3%
	Nuclear	Hrasnu *	0	20	40	70	150	3%
K-Ras	Cytoplasmic	Krascy	0	27	53	85	162	4%
	Nuclear	Krasnu	0	8	23	50	130	4%
N-Ras	Cytoplasmic	Nrascy *	28	113	146	180	250	4%
	Nuclear	Nrasnu *	5	80	100	117	200	4%

Figure 5.3: Examples of distribution of biomarkers in RAS family (Krascy top panel and Nrasnu bottom panel)





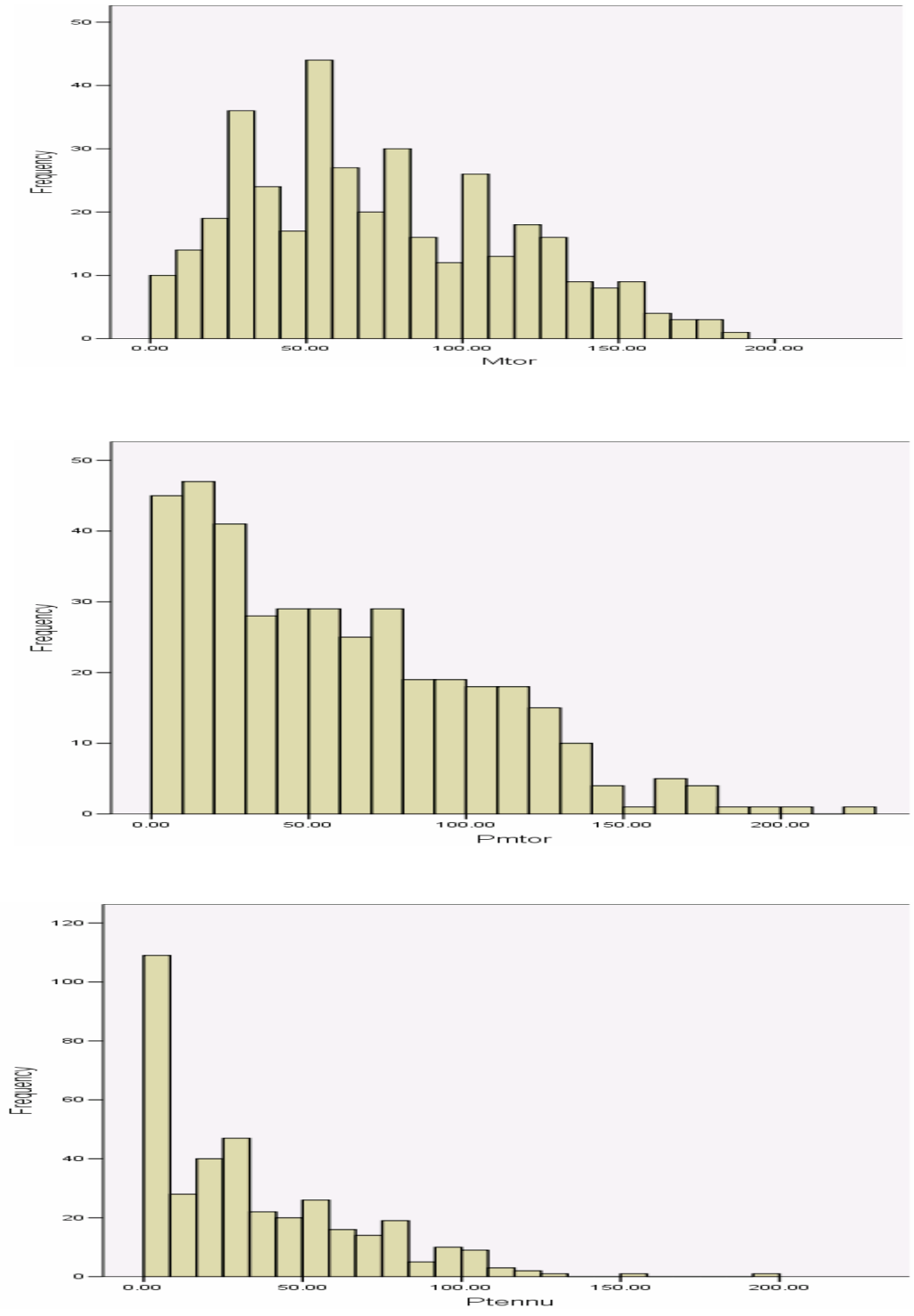
5.3.4 MTOR family

The MTOR family comprised 5 biomarkers. Missing value rate varied from 3% to 15% (Table 5.5). In total, 304 patients (76%) had data available on all 5 biomarkers. The distribution of Mtor was not far from normal but distributions for the other 4 biomarkers were positively skewed. Examples are given in Figure 5.4.

Table 5.5: Distributional statistics and rate of missing values for MTOR family set

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
mTOR	Cytoplasmic	Mtor	0	40	65	105	190	6%
Phospho mTOR	Cytoplasmic	Pmtor *	0	20	50	70	90	3%
Phospho p70S6K (398)	Cytoplasmic	Pp70s6k3 *	0	0	13	85	43	15%
PTEN	Cytoplasmic	Ptency *	0	13	33	50	63	15%
	Nuclear	Ptennu *	0	5	25	53	200	7%

Figure 5.4: Examples of distribution of biomarkers in MTOR family (Mtor top panel, Pmtor middle panel, and Ptennu bottom panel)



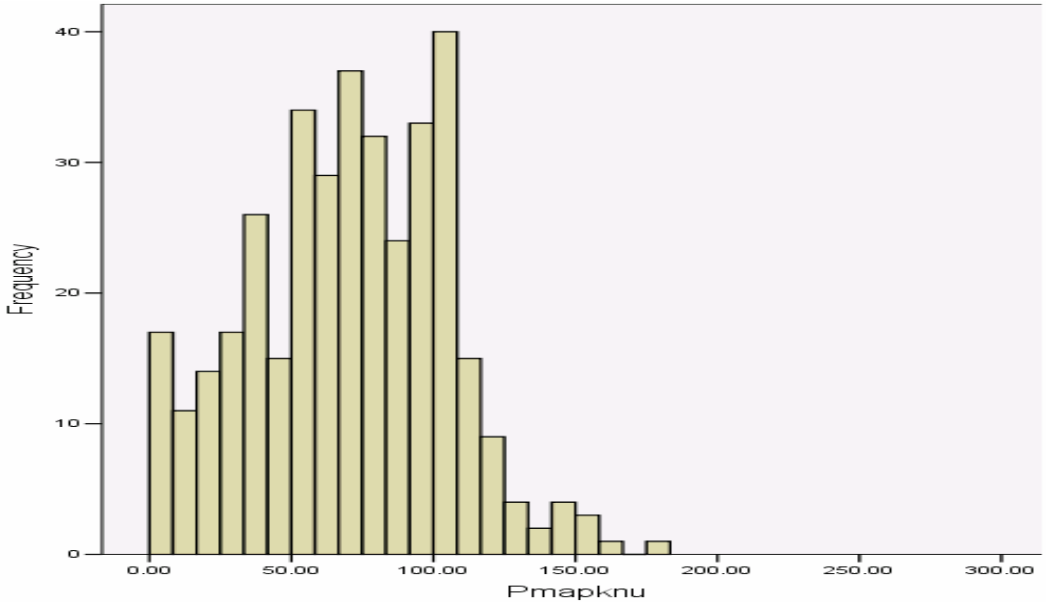
5.3.5 MAPK family

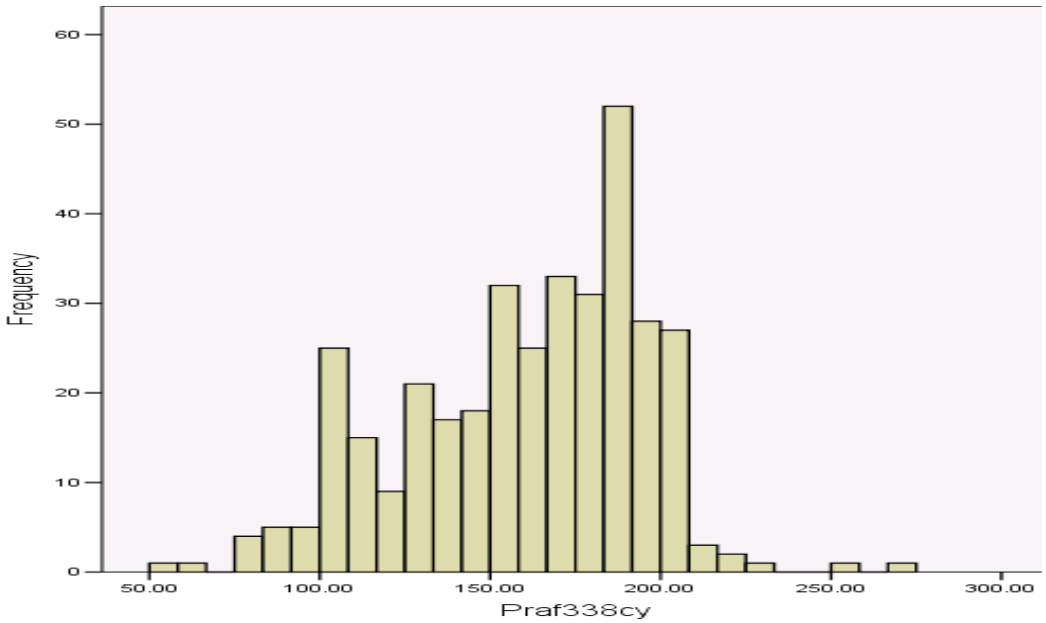
The MAPK family comprised 10 biomarkers. Raf1cy and -nu and Praf338cy and -nu had the highest missing rate at about 11% missing rate (Table 5.6). Distributions of Pmapkcy and -nu and Praf259cy and -nu were skewed while rest of biomarkers exhibited approximately normal distributions. Some examples are given in Figure 5.5.

Table 5.6: Distributional statistics and rate of missing values for MAPK family set

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
Mapk p42/44	Cytoplasmic	Mapkcy	0	70	110	147	260	6%
	Nuclear	Mapknu	0	57	80	107	180	6%
Phospho MAPK IHC	Cytoplasmic	Pmapkcy *	0	10	40	70	185	8%
	Nuclear	Pmapknu *	0	45	72	95	180	8%
Phospho Rat (ser 259)	Cytoplasmic	Praf259cyy	0	23	70	127	280	9%
	Nuclear	Praf256nu	0	0	2	10	93	9%
Phospho Rat (ser 338)	Cytoplasmic	Praf338cy *	58	136	167	190	275	11%
	Nuclear	Praf338nu *	5	113	135	158	220	11%
Raf-1	Cytoplasmic	Raf1cy	0	83	123	153	280	12%
	Nuclear	Raf1nu	0	92	108	123	200	12%

Figure 5.5: Examples of distribution of biomarkers in MAPK family (Pmapknu top panel and Praf338cy bottom panel)





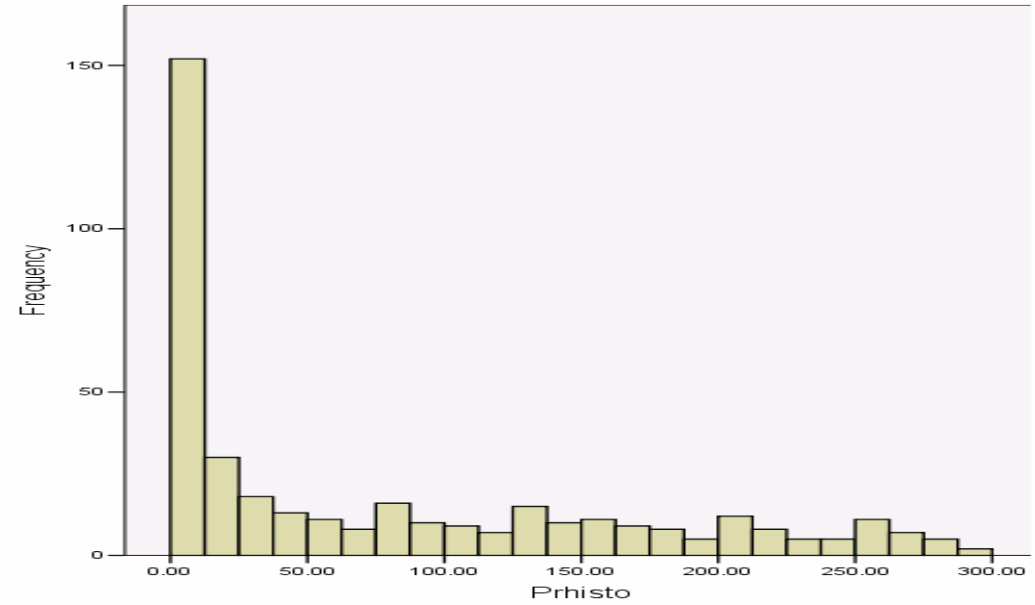
5.3.6 PgR family

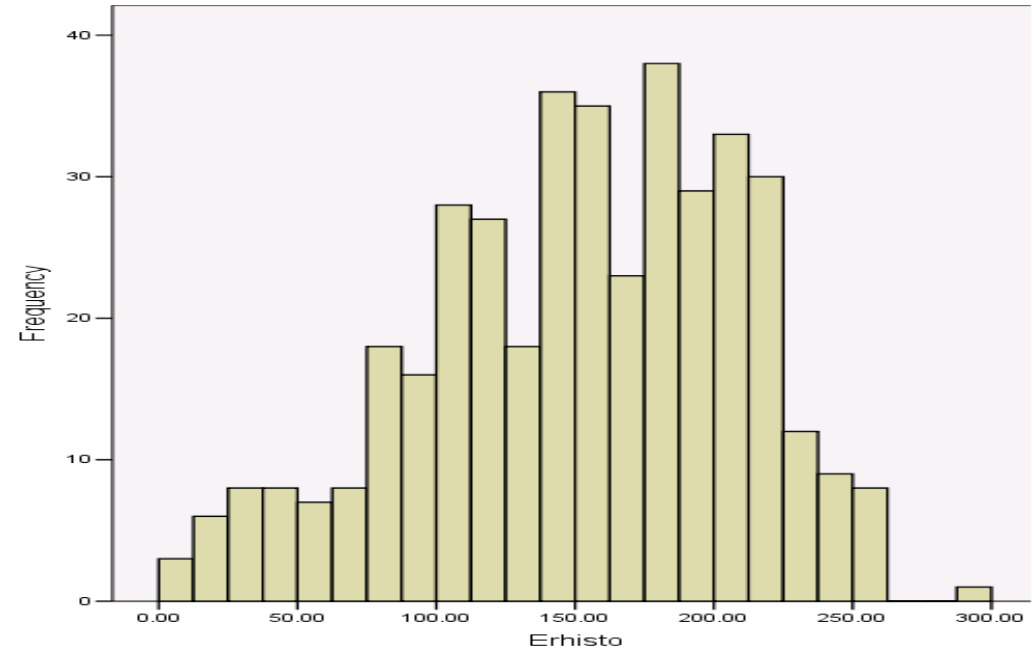
PgR family comprised 11 biomarkers, of which 3 had no missing data (Table 5.7). The missing value rate for the rest of biomarkers was less than 10%. The distributions of the majority of biomarkers were skewed. Some examples are given in Figure 5.6.

Table 5.7: Distributional statistics and rate of missing values for PgR family set

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
Estrogen receptor	Cytoplasmic	Ercy *	0	15	50	80	150	0%
	Nuclear	Erhisto	10	112	153	197	300	0%
Estrogen receptor beta	Cytoplasmic	Erbcy *	0	0	50	100	300	9%
	Nuclear	Erbnu	0	100	125	163	275	9%
Progesterone receptor	Nuclear	Prhisto *	0	0	35	140	300	4%
Phospho 118 ER	Cytoplasmic	P118cy *	0	100	175	225	300	5%
	Nuclear	P118nu	0	113	145	175	270	5%
	Membrane	P118me *	0	0	0	0	133	0%
Phospho 167 ER	Cytoplasmic	P167cy *	0	0	50	100	250	10%
	Nuclear	P167me *	0	0	0	0	165	9%
	Membrane	P167nu	0	63	100	140	250	5%

Figure 5.6: Examples of distribution of biomarkers in PgR family (Prhisto top panel and Erhisto bottom panel)





5.3.7 HER family

Eighteen biomarkers formed the HER family. One biomarker had more than 40% missing rate (Table 5.8). In comparison with other families, majority of HER biomarkers had a fairly high missing rate. Figure 5.7 gives the distribution of HER2.

Figure 5.7: Example of distribution of a biomarker in HER family (Her2)

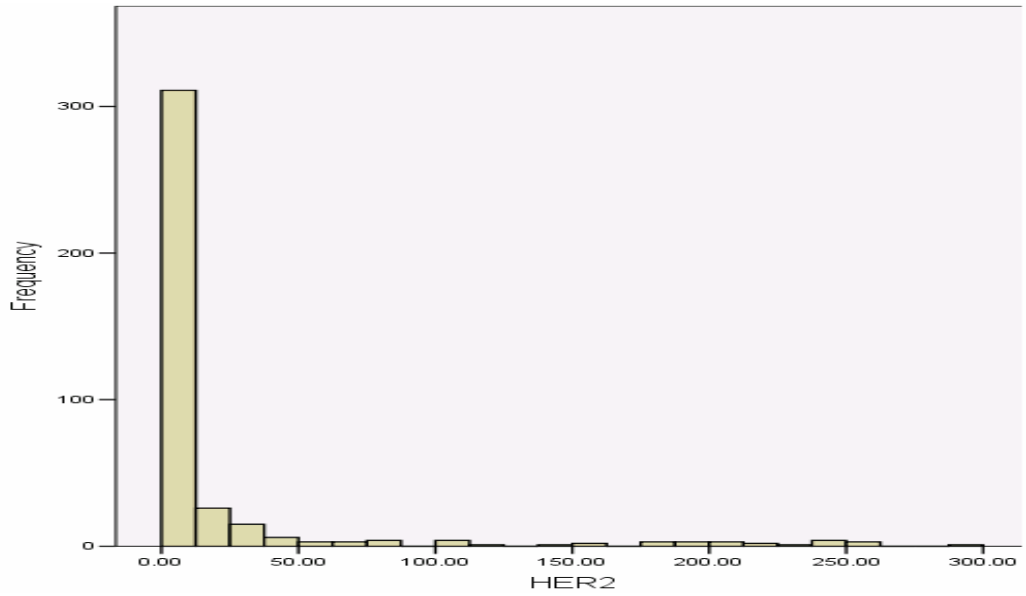


Table 5.8: Distributional statistics and rate of missing values for HER family set

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
Phospho HER2	Cytoplasmic	Pher2cy *	0	60	100	143	300	6%
	Nuclear	Pher2nu *	0	25	43	65	100	6%
	Membrane	Pher2me *	0	60	100	143	300	6%
HER2	Membrane	Her2 *	0	0	0	7	292	1%
HER2 FISH gene/ chromosome 17 ratio		Her2fish *	0.9	1.1	1.1	1.2	8	12%
AIB1	Nuclear	Aib1 *	0	30	58	100	205	6%
JR HER3	Cytoplasmic	Jrh3cy *	0	50	83	140	265	12%
	Nuclear	Jrh3nu *	0	30	58	90	250	12%
	Membrane	Jrh3me *	0	0	0	0	180	41%
HER4 HFR1	Cytoplasmic	H4hfr1cy *	0	26	75	125	253	11%
	Nuclear	H4hfr1nu *	0	35	63	93	200	11%
	Membrane	H4hfr1me *	0	0	0	20	150	11%
HER4 H4.77.16	Cytoplasmic	H4jrcy *	0	0	37	79	250	15%
	Nuclear	H4jrnu *	0	0	0	25	200	15%
	Membrane	H4jrme *	0	0	0	23	200	15%
Categorical variables			**value (frequency)**					
		Egfrmax	0 (386)	1 (1)	2 (2)	3 (3)		2%
		Aib1fis1	0 (337)	24 (1)				10%
		Aib1fis2	0 (334)	9 (1)	18 (2)			10%

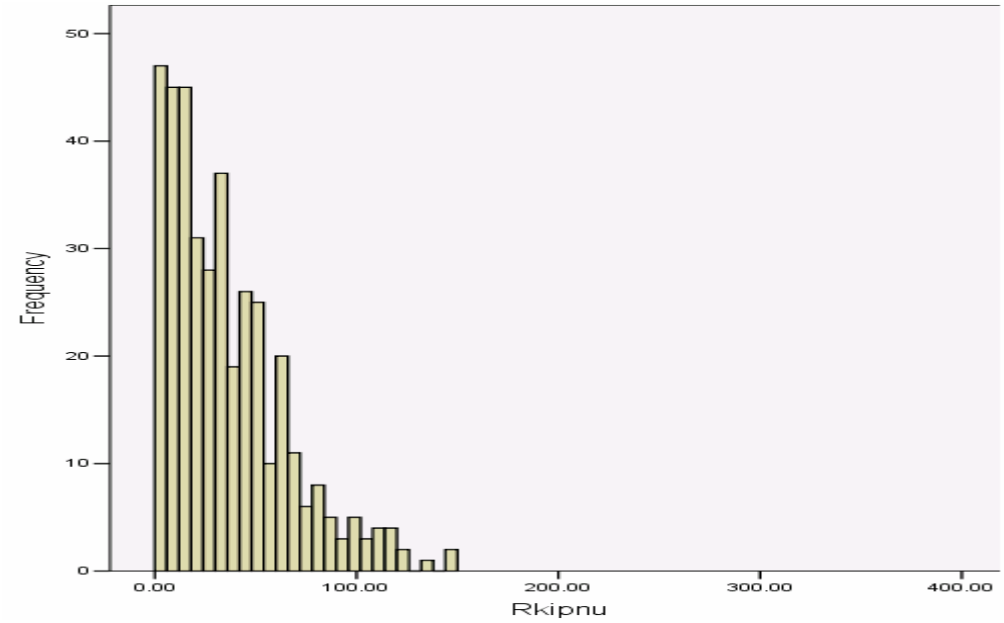
5.3.8 'Non-family' biomarkers

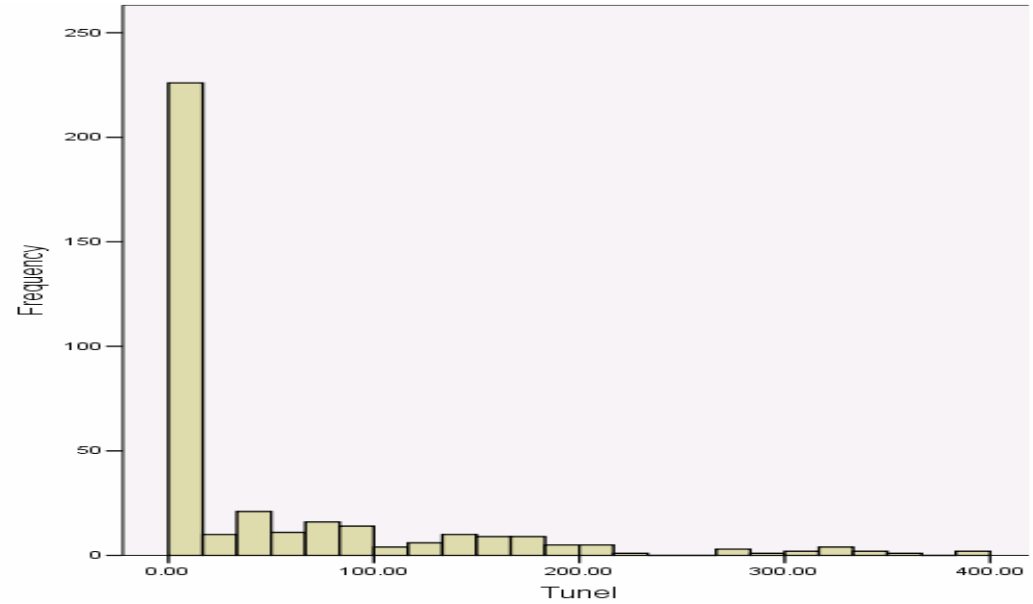
Seven biomarkers were not assigned to any of the biomarker families. The rate of missing value was less than 10% for all biomarkers (Table 5.9). The distributions of Tescy and -nu were approximately normal while the remaining biomarkers exhibited skewed distributions (Figure 5.8).

Table 5.9: Distributional statistics and rate of missing values for non-family biomarkers

Biomarker	Expression	Abbreviation in thesis	Min	Q1	Q2	Q3	Max	Missing (%)
Tace	Cytoplasmic	Tace *	0	3	10	20	135	5%
Tacep	Cytoplasmic	Tacep *	0	3	10	20	90	5%
Tunel	-----	Tunel *	0	0	0	72	400	10%
TES	Cytoplasmic	Tescy	0	75	112	170	275	7%
	Nuclear	Tesnu	0	87	117	150	223	7%
rKIP	Cytoplasmic	Rkipcy *	0	47	85	117	230	7%
	Nuclear	Rkipnu *	0	12	28	50	150	3%

Figure 5.8: Examples of distribution of biomarkers in non-family set (Rkipnu top panel and Tunel bottom panel)





5.3.9 Clinical variables

Distributional statistics for tumour size are given. In addition, frequency distribution of patients based on tumour grade and stage is presented.

Table 5.10: Distributional statistics and rate of missing values for clinical variables

Tumour size (cm)		Stage		Grade	
Statistic	Value	Value	Frequency	Value	Frequency
Min	0.2	1	192	1	99
Q1	1.5	2	107	2	192
Q2	2	3	69	3	99
Q3	3				
Max	11				
Miss	6%	Miss	8%	Miss	3%

5.3.10 Pattern of missing data

The number of biomarkers in each set with number (percentage) of patients with available data on all variables within that set is reported in Table 5.11. Only 51% of patients had complete data for HER family biomarkers whereas approximately 90% of patients had data available on all biomarkers in each of RAS and AKT family. Only 126 patients (31%) had complete data on all 72 biomarkers and 3 clinical variables.

The biomarkers with more than 10% missing value are listed in Table 5.12. They belonged to just 3 families: HER, BAD, and MAPK. The highest missing rate 41% but the remaining had missing value between 11% and 15.5%.

Indeed all 5 biomarkers of the BAD family had more than 10% missing value.

Table 5.11: Number of patients with available data in family sets of biomarkers

Family/ set	Number of variables	Number (percentage) of patients with data available on all variables
RAS	6	372 (93%)
AKT	10	356 (89%)
PgR	11	334 (83%)
MAPK	10	325 (81%)
Non-family set	7	326 (81%)
MTOR	5	304 (76%)
BAD	5	294 (73%)
HER	16	203 (51%)

Table 5.12: Biomarkers with more than 10% missing value

Family	Variable	Number (percentage) of patients with missing value
HER	Jrh3me	166 (41%)
	H4jrcy	61 (15%)
	H4jrmem	61 (15%)
	H4jrnu	61 (15%)
	Jrh3cy	49 (12%)
	Jrh3nu	49 (12%)
	Her2fish	47 (12%)
	H4hfr1me	43 (11%)
	H4hfr1nu	43 (11%)
	H4hfr1cy	43 (11%)
BAD	Pp70s6k3	62 (15%)
	Pbad112c	56 (14%)
	Baxcy	53 (13%)
	Badcy	49 (12%)
	Bclxl	45 (11%)
MAPK	Raf1cy	47 (12%)
	Raf1nu	46 (11%)
	Praf338cy	44 (11%)
	Praf338nu	44 (11%)

5.4 Summary

The distribution of majority of biomarkers was far from being normal. However, some of the distributions were hugely skewed (for examples see HER and PgR families) in which histoscore value for 50% or even 75% of patients were zero. This indicates that those biomarkers might not be helpful for modelling unless a very large threshold effect exists.

In addition presence of outliers in biomarkers, for example see histogram of Ptennu, in the case such biomarkers predict the outcome, might hugely depend to the data. Small changes in the data might affect predictive ability of such biomarkers.

Chapter 6 NOTTINGHAM PROGNOSTIC INDEX FOR BREAST CANCER

6.1 Introduction

In later chapters I will need to compare the biomarker models I developed with the Nottingham Prognostic Index (NPI). Details of the original development of the NPI [Haybittle JL et al., 1982] have been given in Chapter 2.

Comparison of performance of models requires similar approach to categorise patients into the risk groups. The standard NPI categorised patients into 3 risk groups. However, for the biomarker models developed, patients will be categorised into 4 risk groups by applying as cut offs the quartiles of the distribution of the risk score (see section 4.4.5). Therefore, to perform a fairer comparison, and to check the

effect of stratification into 4 groups relative to 3, I need to categorise patients into 4 risk groups based on NPI risk scores.

However, a further issue is that, in comparison with NPI, biomarker models are optimised for the data set used in this research, since by definition they give the best fit to the data. Furthermore, the data set used to develop biomarker models and to quantify the performance are the same. The Glasgow data set is the training sample for biomarker models but the validation sample for NPI. Therefore to ensure the fairest possible NPI comparator, I recalculated the new NPI risk scores using the current data set and then categorised patients into 4 risk groups using new risk scores.

6.2 Aims

The aims of this part of the research are to:

1. Categorise patients into 4 risk groups on the basis on NPI risk scores and check effect of categorisation to 4 relative to 3 groups on risk stratification

2. Refit NPI using the current data set as training set, and to check whether in comparison with standard NPI, it gives greater discrimination between risk groups

6.3 Methods

6.3.1 Calculation of standard NPI and categorisation of patients into 3 risk groups

Multivariable Imputations by Chained Equations (MICE) [Van Buuren S et al., 1999; Van Buuren S and Oudshoorn K, 2000] method was applied to create 10 imputed data sets (section 4.4.4). In each data set, standard NPI (0.2 x size (cm) + stage + grade) was calculated [Haybittle JL et al., 1982]. For each patient, the final NPI score was calculated by averaging her scores across the 10 imputed data sets. Using final scores, standard cut offs were applied at 3.4 and 5.4 to categorise patients into 3 risk groups (NPI^{std3}) (Table 6.1 row 1).

6.3.2 Calculation of standard NPI and categorisation of patients into 4 risk groups

To categorise the patients into 4 risk groups, cut offs were applied at quartiles of the NPI derived from imputed data sets (NPI^{q4}) (Table 6.1 row 2). As an alternative, I applied the published cut offs (2.4, 3.4, and 5.4), that is with one lower cut off to subdivide the lowest risk group into two (NPI^{std4}) (Table 6.1 row 3) [Galea MH et al., 1992].

6.3.3 Recalculation of NPI using the current data set (recNPI)

For each of 10 imputed data sets, a risk score was recalculated by multiplying nodal status, grade, and tumour size values for each patient to the estimated specific regression coefficient corresponding to that data set. The average of recalculated scores across the 10 imputed data sets was used as final index (recNPI). Using recalculated risk scores patients were assigned into 4 risk groups applying the cut offs at quartiles ($recNPI^{q4}$) (see Table 6.1 row 4).

I also compared standard NPI with published cut offs with recalculated NPI. Therefore, I applied cut offs to recalculated risk scores so as to create risk group containing similar patients to that of NPI^{std4} ($recNPI^{std4}$) (see Table 6.1 row 5).

All models were compared as explained in section 4.4.7. Plotting Kaplan Meier (K-M) curves, the numbers of patients at start and followed for 3, 5, 7, and 10 years is reported below each plot.

Table 6.1: Description of NPIs calculated and cut offs applied to assign patients into risk groups

Row	Label	Index calculation	Risk groups	Cut offs
1	NPI^{std3}	Standard formula	Standard 3 risk group	(3.4, 5.4)
2	NPI^{q4}	Standard formula	4 equal sized risk group	quartiles
3	NPI^{std4}	Standard formula	Published 4 risk groups	2.4, 3.4, 5.4
4	$recNPI^{q4}$	Recalculated formula	4 equal sized risk group	quartiles
5	$recNPI^{std4}$	Recalculated formula	Mirror of NPI^{std4}	

6.4 Results

6.4.1 Standard NPI with 3 risk groups

The numbers of patients with missing values on node, grade, and tumour size were 33, 11, and 22 respectively. Nodal status had the highest missing rate (about 8%). In total, 343 patients (86%) had data available on all 3 variables of which, 88 had experienced recurrence.

Discrimination (C-index) and predictive ability of index was 72% and 14% respectively. Estimated Recurrence Free Survival (RFS) rates, in the lowest and highest risk groups are given in Table 6.2. Estimated 7-year RFS rate in the lowest risk group was 89% which was 6 percentage points smaller than the target (95%). K-M survival curves are given in Figure 6.1 (left panel).

6.4.2 Standard NPI with 4 risk groups

The conventional cut offs for NPI create 3 risk groups which, in the current data set, resulted in a large low risk group (n=133, 33% of patients) and a smaller high risk group (69, 17%). Applying quartiles as cut offs (3.3, 4.24, and 4.8), patients were categorised into 4 equal size risk groups (Table 6.2). Creation of 4 risk groups rather than 3 gave a smaller low risk group (102 versus 133) containing fewer number of recurrences (10 versus 15). However, this had a marginal impact on estimated RFS rates in the lowest-risk group (91% for 4 and 89% for 3 risk groups). K-M curves comparing these risk stratifications are given in Figure 6.1 (left and middle panels).

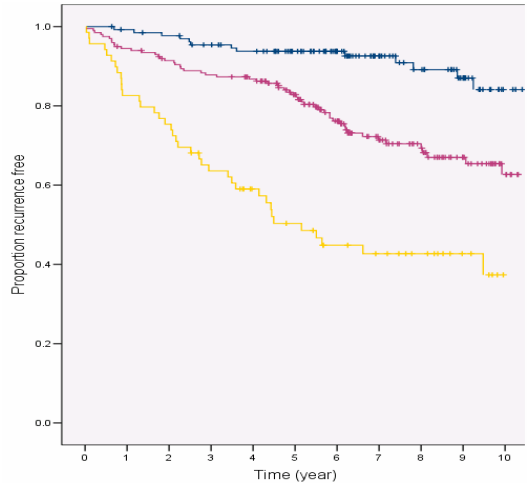
I then created 4 risk groups applying published cut offs at 2.4, 3.4, and 5.4 (NPI^{std4}). This resulted into a very small risk group (n=43, 11% of patients) and very big higher-intermediate risk group (n=212, 53%) (Table 6.2). In comparison with 4 equal size risk grouping, estimated RFS rates of lowest-risk group was increased to 95%, from 91%. In the lowest risk group of NPI^{std4} only 2 early recurrences at 2.26 and 2.52 years of follow-up were observed. That is why RFS rate at 5 years remained constant up to 10th year of follow-up. The biological collaborators for this research hoped to detect a low risk group with 7-year RFS of 95%, which was achieved by the published 4 risk groups of NPI. However, only about one-tenth of patients were categorised into the low risk group. K-M survival curves are given in Figure 6.1 (right panel).

Table 6.2: RFS rates in the lowest and highest risk groups of NPI: standard 3 versus 4-level categorisations

Risk group	Index	N at stat	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	NPI^{std3}	133	94% (90%, 98%)	**89% (83%, 95%)**	79% (65%, 93%)
	NPI^{q4}	102	95% (91%, 99%)	**91% (85%, 97%)**	84% (72%, 96%)
	NPI^{std4}	43	95% (89%, 100%)	**95% (89%, 100%)**	95% (89%, 100%)
Highest	NPI^{std3}	69	44% (32%, 56%)	**42% (30%, 54%)**	36% (20%, 52%)
	NPI^{q4}	100	54% (44%, 64%)	**49% (39%, 59%)**	41% (27%, 55%)
	NPI^{std4}	72	47% (35%, 59%)	**45% (33%, 57%)**	39% (23%, 55%)
PSEP for NPI^{std3} risk groups			50%	**47%**	43%
PSEP for NPI^{q4} risk groups			41%	**42%**	43%
PSEP for NPI^{std4} risk groups			48%	**50%**	56%

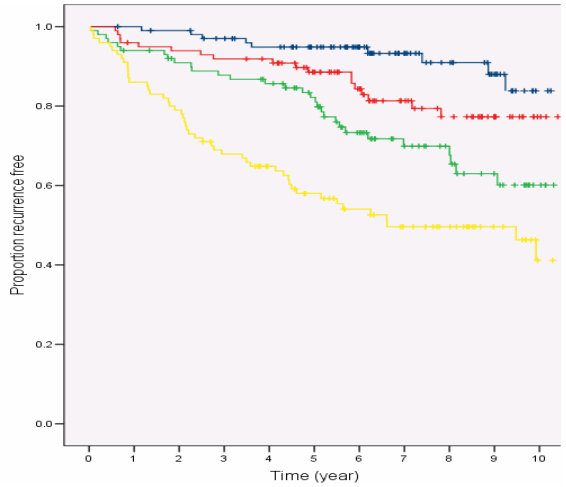
Figure 6.1: K-M curves for standard NPI: traditional 3 risk groups (left panel) versus 4 risk group schemes (middle and right panels)

NPI^{std3}



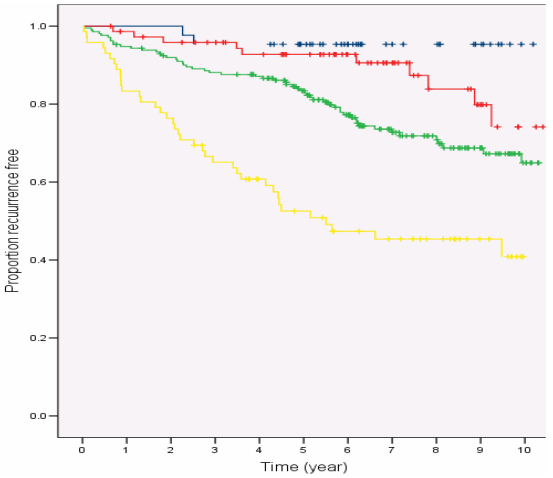
133	121	103	63	38	19
199	171	138	79	42	23
69	42	28	19	9	3

NPI^{q4}



102	94	80	47	27	13
99	90	73	45	24	13
100	84	69	38	22	12
100	66	47	31	16	7

NPI^{std4}



43	41	35	19	15	5
74	65	55	34	18	10
212	83	148	86	48	27
72	45	31	22	11	3

6.4.3 Recalculation of NPI (recNPI)

Recalculating the index, a slight improvement in the discrimination (C-index 73.5% versus 72%) and predictive ability (R-square 16% versus 14%) was seen. However, using quartile risk groups, no noticeable difference to estimated RFS rates was seen for recalculated and standard NPI (Table 6.3). K-M curves are given in Figure 6.2.

Table 6.3: Estimated RFS rates in the lowest and highest risk groups of 4 risk group stratifications based on standard and recalculated NPI

Risk group	Index	N at stat	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	$recNPI^{q4}$	101	94% (90%, 98%)	**90% (82%, 98%)**	77% (61%, 93%)
	NPI^{q4}	102	95% (91%, 99%)	**91% (85%, 97%)**	84% (72%, 96%)
Highest	$recNPI^{q4}$	98	52% (42%, 62%)	**50% (40%, 60%)**	41% (27%, 55%)
	NPI^{q4}	100	54% (44%, 64%)	**49% (39%, 59%)**	41% (27%, 55%)
PSEP for $recNPI^{q4}$ risk groups			42%	**40%**	36%
PSEP for NPI^{q4} risk groups			41%	**42%**	43%

Applying cut offs at quartiles of standard and recalculated NPI, the distribution of patients into risk groups was such that risk groups derived from recalculated index, relative to standard index, classified 13 recurred patients into a more appropriate risk group and the same number into a less appropriate risk group, giving a net gain of zero. Corresponding figures for non-recurred patients were 31 and 27 respectively. This gave a net gain of 1.4 percentage points which was not significant (P=0.30).

Thus in this data set recalculation of index, and reclassification of patients, did not make a significant difference to discrimination in recurrence across risk groups.

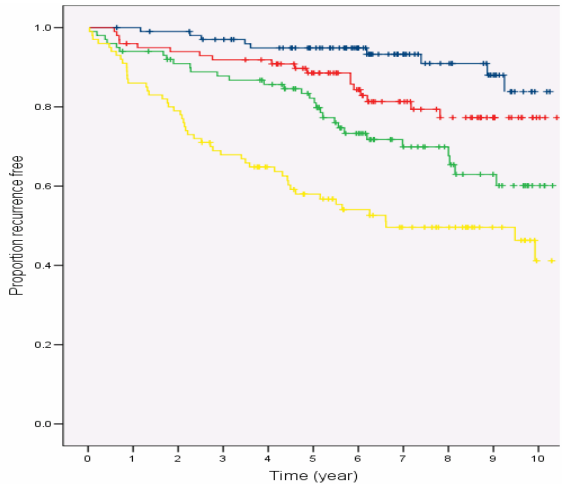
I also created 4 risk groups with unequal number of patients by applying the cut offs to create 4 risk groups containing 43, 74, 212, and 72 patients (the same as risk groups derived by applying published cut offs at standard NPI (NPI^{std4})). Cut offs applied were 1.35, 1.71, and 2.8. Estimated RFS rates were fairly similar (Table 6.4). K-M survival curves are given in Figure 6.2.

Table 6.4: Estimated RFS rates in the lowest and highest risk groups of two 4 risk group stratifications based on recalculated NPI

Risk group	Index	N at stat	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	$recNPI^{std4}$	43	95% (89%, 100%)	**95% (89%, 100%)**	95% (89%, 100%)
	NPI^{std4}	43	95% (89%, 100%)	**95% (89%, 100%)**	95% (89%, 100%)
Highest	$recNPI^{std4}$	72	43% (31%, 55%)	**43% (31%, 55%)**	36% (20%, 52%)
	NPI^{std4}	72	47% (35%, 59%)	**45% (33%, 57%)**	39% (23%, 55%)
PSEP for $recNPI^{std4}$ risk groups			52%	**52%**	49%
PSEP for NPI^{std4} risk groups			48%	**50%**	56%

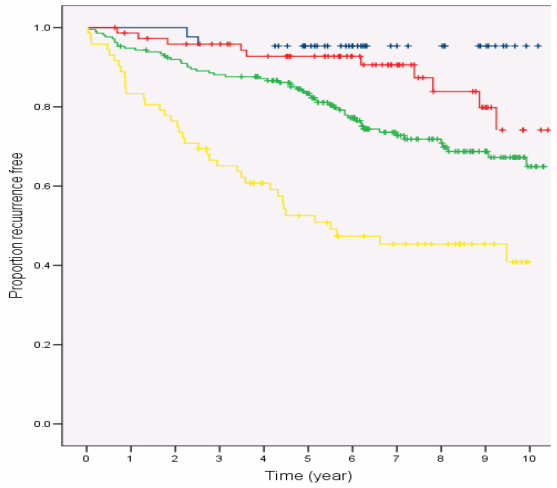
Figure 6.2: K-M survival curves for standard (top panels) and recalculated NPI (bottom panels)

NPI^{q4}



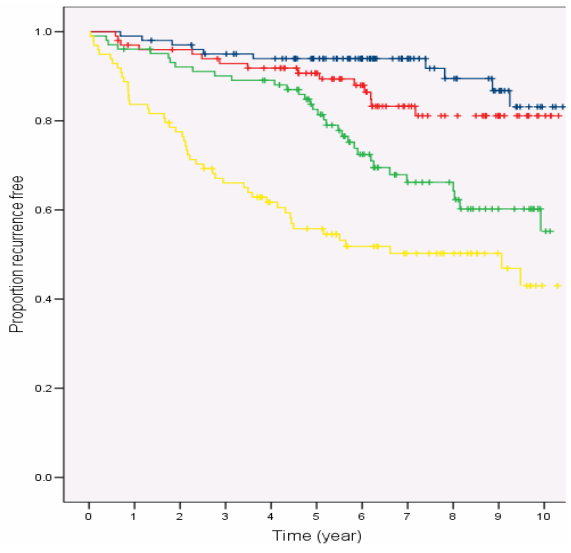
102	94	80	47	27	13
99	90	73	45	24	13
100	84	69	38	22	12
100	66	47	31	16	7

NPI^{std4}



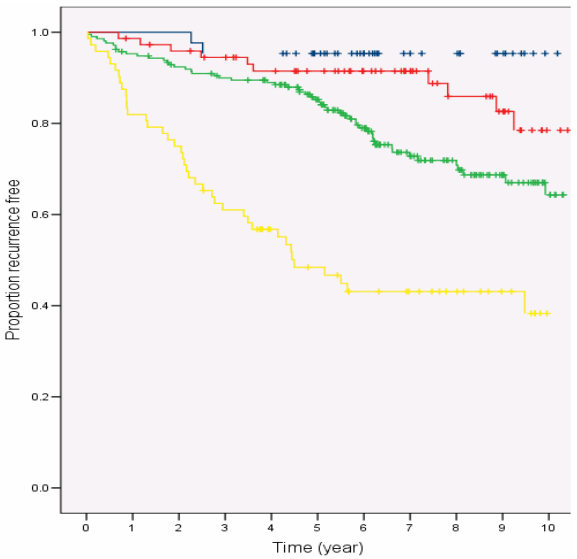
43	41	35	19	15	5
74	65	55	34	18	10
212	83	148	86	48	27
72	45	31	22	11	3

recNPI^{q4}



101	93	79	50	30	16
100	89	72	42	23	12
102	90	72	39	21	11
98	62	46	30	15	6

recNPI^{std4}



43	41	35	19	12	5
74	66	56	39	24	13
212	184	150	84	43	24
72	43	28	19	10	3

6.5 Discussion

6.5.1 Creation of 4 groups instead of 3

As mentioned in Chapter 4, in development of biomarker models cut offs will be applied at quartiles to assign patients into 4 risk groups. To find a fairer comparator to biomarker models, I did the same for NPI risk scores. Applying the cut offs at quartiles, in comparison with standard 3 risk groups, no noticeable difference in RFS rate of lowest-risk group was seen. On the other hand, once published 4-group cut offs were applied (2.4, 3.4, and 5.4), number of patients formed the lowest risk group were 43 including 2 early recurrences. Therefore RFS rates were slightly improved. However, only 11% of patients fell into the lowest risk group.

6.5.2 Recalculation of index

I refitted the NPI and categorised the patients into 4 groups. However, I did not have information on all 9 variables used in calculation of NPI. I therefore only used the 3 clinical variables in a Cox regression model. A slight improvement in C-index and R-square was seen. However, in terms of risk stratification, K-M curves were comparable.

This was consistent with results of other studies advocating standard use of NPI [Okugawa H et al., 2005; D'Eredita' G et al., 2001]. The advantage of application of standard index and risk groups is that results of different studies will be comparable.

6.5.3 Ability to detect low-risk patients

Results summarised in Chapter 2 (Table 2.2) suggested that NPI was not able to detect a specific lowest-risk group with sufficiently low risk of recurrence. However, it has been commented that 'a subgroup of women within the good prognostic group with NPI scores ≤2.4 has an excellent prognosis with a 15-year survival of 94%, thus representing a group of patients potentially cured by locoregional treatment alone' [Kollias J et al., 1999].

Any patient with nodal status and grade score of 1 and tumour size of ≤ 2 cm, have NPI risk score ≤ 2.4. Although application of split at 2.4 can detect patients with excellent prognosis, only a small proportion of patients might meet these criteria (43 patients (11%) in the Glasgow data set).

Results presented in this chapter showed that while, at 5 years, the short-term recurrence free rate of the lowest-risk group defined by NPI was around 95%, a gentle decreasing trend was seen in K-M curves after fifth year of follow up. As an example, actuarial 5 and 10-year RFS rates in the lowest risk group of standard NPI were 94% and 79% respectively. Only for best prognosis risk group (with NPI≤ 2.4) were short-term RFS at 5 years and long-term RFS at 10 years the same, because no recurrences were observed after fifth year.

6.6 Overview

Results presented showed that recalculation of NPI did not improve risk prediction and therefore I am reassured with respect to using standard NPI as the comparator for biomarkers which I will develop. Furthermore, in terms of estimates RFS rates in the risk groups, there was a slight difference between standard 3 and 4 equal risk group schemes. In development of biomarker models, I will use quartiles for risk grouping. Therefore, to estimate Net Reclassification Index (NR Index), and to compare RFS rates in the lowest and highest risk groups, I will use NPI with 4 equal risk groups (NPI^{q4}) as the basic risk grouping scheme.

6.7 Chapter summary

- Creation of 4 equal sized risk groups did not improve ability of NPI to detect low-risk patients. Applying standard cut offs at 2.4, 3.4, and 5.4, the goal of detection of a subset of patients with 7-year RFS of 95% was achieved. However, small proportion of patients had NPI as low as 2.4.

- Comparison of recalculation and standard NPI risk groupings, no noticeable difference in estimated RFS rates was seen. On the other hand, use of standard NPI ensures results of different studies are comparable. Therefore, standard NPI with cut offs at quartiles will be used as comparator in latter chapters.

Chapter 7 SCREENING AND UNIVARIATE FUNCTIONAL FORM OF ASSOCIATION FOR BIOMARKERS

7.1 Introduction

The majority of biomarkers in the data set exhibited a skewed distribution (see Chapter 5 for details). When data is skewed, it is common in prognostic modelling to apply a pre-specified transformation such as logarithmic, prior to analysis, to make a linearity assumption plausible (even if not optimal). An alternative method frequently used is to dichotomise the continuous variables so as to simplify the analysis.

However, the answer to the question ‘Is there an effect?’ depends to a great extent on the choice of risk function [Hollander N and Schumacher M, 2006]. Therefore, appropriate methods should be applied to reveal the optimum form of risk function for continuous variables. This raises the issue of how to find the optimum form of association for each biomarker.

When the ability of different statistical techniques to detect the right form of risk function for continuous variable have been compared, it has been seen that Fractional Polynomial (FP) is the best technique to deal with ‘linear and polynomial’ effects. FP was also a good approximation for threshold or V shape effects [Hollander N and Schumacher M, 2006] (see Chapter 3 section 3.4.4 for details). Furthermore, and importantly, FP does not inflate type one error [Ambler G and Royston P, 2001].

I wished to apply a range of different procedures to all 72 biomarkers, in order to compare the performance of alternative statistical methods in terms of detection of form of association and also to select a set of univariately potential informative biomarkers. Biomarkers associated with Recurrence Free Survival (RFS) with P-value <0.10 (corrected for multiple testing) were considered to be univariately informative. From here on in this thesis, these biomarkers are denoted as univariately informative biomarkers.

7.2 Aims

The Event Per Variable rule [Peduzzi P et al., 1995] would require over 720 recurrences to deal with 72 biological variables available in the data set. However, only 112 recurrences were observed out of 401 patients recruited. The screening process applied in this chapter lays the foundation for the methods chosen to be applied in the later chapters. The main aims of this part of the research are as to:

1. Apply a range of screening methods to detect univariate form of association for biomarkers and to compare selection of biomarkers by different methods

2. Select a reduced set of potentially informative biomarkers based on univariate analysis of association with outcome, to be used in the next chapters in development of the multifactorial models

7.3 Methods

7.3.1 Detection of form of association

Out of 72 biomarkers only 3 ones were categorical (see Chapter 5, Table 5.8 for details). Five methods were applied to all 69 continuous biomarkers so as to detect a range of functional forms of associations: 2 methods to detect linear or polynomial effects, 2 methods to detect threshold effects, and 1 method to detect non-ordinal effects (Table 7.1). For each univariate test, patients with missing data on that biomarker were excluded.

i) Detection of linear or polynomial effects

Fractional Polynomial modelling

Two approaches were applied to detect linear or polynomial effects. Firstly, Fractional Polynomial (FP) was applied to each continuous biomarker [Royston P and Altman DG, 1994] (section 4.4.2). FP is a data-driven (data-dependent) technique which works with continuous variables.

Since Sauerbrei et al. recommended the use of FP to detect linear and polynomial effects [Sauerbrei W et al., 2007] and Hollander et al. showed that FP is a good approximation for threshold and V shape effects [Hollander N and Schumacher M, 2006] results of other methods were compared with FP.

Linear Cox model

Secondly, I assume that a pre-specified liner form was adequate. Therefore, variables were kept in the continuous form and linear Cox model (special case of FP1 when power is equal to 1) was applied.

ii) Detection of threshold effects

Two approaches were then applied to dichotomise biomarkers and to detect threshold effects: minimum P-value method and dichotomisation at quartiles:

Minimum P-value method

Optimal split for biomarkers were found applying minimum P-value method (section 4.4.3) [Williams BA et al., 2006]. To correct for multiple testing, biomarkers with P-value < 0.005 were taken as informative (equivalent to 0.10 in linear Cox model) [Altman DG et al., 1994]. This technique is data-driven.

Dichotomisation at quartile(s)

Secondly, all 69 biomarkers were dichotomised, in turn, only at lower quartile (Q1), median (Q2), and upper quartile (Q3) (3 comparisons for each biomarker). This approach is named ‘quartile dichotomisation’. Biomarkers with P-value < 0.033 at any of quartiles were declared as informative. This is a less extreme data-dependent method.

iii) Detection of non-ordinal effects

In order to check for non-ordinal effects, as shown in Figure 7.1, I made 4 comparisons for each biomarker. In each case, patients in the shadowed quartile were compared with unshadowed quartiles. Variables with P-value <0.025 were declared as informative. This approach was named 'non-ordinal quartile dichotomisation'. This is a data-dependent method.

Figure 7.1: Comparisons made to detect non-ordinal effects

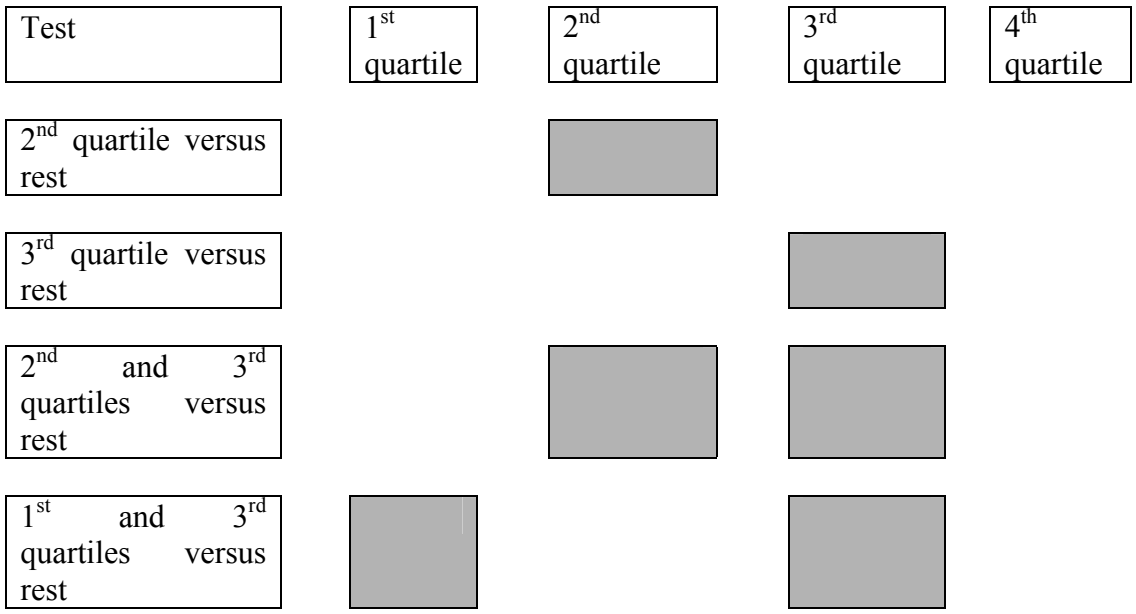


Table 7.1: Screening methods applied to estimate form of risk function of 69 continuous biomarkers

Purpose: to capture effects which are of form	Method	Nature of method	The type of variable will be	Number of tests applied (per biomarker)	Threshold P-value used to declare biomarker is informative
Linear or Polynomial	FP	Data driven	Continuous	36 FP2 plus 8 FP1	0.10
Linear	Linear Cox	Pre-specified	Continuous	1	0.10
Threshold	Minimum P-value	Data driven	Binary	All values excluding the outer 20% in the distribution	0.005
Threshold	Quartile dichotomisation	Pre-specified but partly data driven	Binary	3	0.033
Non-ordinal	Non-ordinal quartile dichotomisation	Data dependent	Binary	4	0.025

7.3.2 Selection of informative biomarkers and their form

In order to select a reduced set of informative biomarkers, I restricted methods applied to only 3 screening methods, one for each of 3 associations described above, which optimised the form or place of split and predicted RFS at a 0.10 significance level: Fractional Polynomial (FP) to select 'linear or polynomial' effects at 0.10 level (section 4.4.2), minimum P-value to select threshold effects at 0.005 level (section 4.4.3), and non-ordinal quartile dichotomisation to select non-ordinal effects at 0.025 level.

When FP and another screening method selected a biomarker to be informative, the form which was indicated by FP was taken to apply. That is because FP does not dichotomise the data and avoids loss of information [Royston P et al., 2006]. In the case of overlap between minimum P-value and 'non-ordinal quartile dichotomisation' the more complex association (non-ordinal) was taken to apply.

To check the informativeness of 3 categorical biomarkers (see Table 5.8), patients with non-zero values were grouped together thus creating 3 binary variables. This is because small number of patients had a non-zero histoscore value. A univariate binary Cox model was applied and those significant at a 0.10 level were selected.

7.4 Results

7.4.1 Application of methods to detect form and compare selection of biomarkers

i) Application of FP method

Applying FP2 method, 12 biomarkers were identified as informative (Table 7.2) and for 3 of these biomarkers the association was polynomial. The shapes of the risk function for biomarkers with polynomial effects are plotted in Figure 7.2. For two of them (Krascy and Rkipnu) the best functional form was expressed by FP2 (which means two powers are required to explain their effect). Corresponding powers were (3, 3) for Krascy and (1, 0.5) for Rkipnu respectively. In addition, a reciprocal square transformation (an FP1 with power (-2)) was necessary to capture the association of Ptennu with RFS.

ii) Application of linear Cox model

Applying univariate Cox model, a total of 9 variables were selected as informative. Polynomial associations detected by FP method were missed.

iii) Application of minimum P-value method

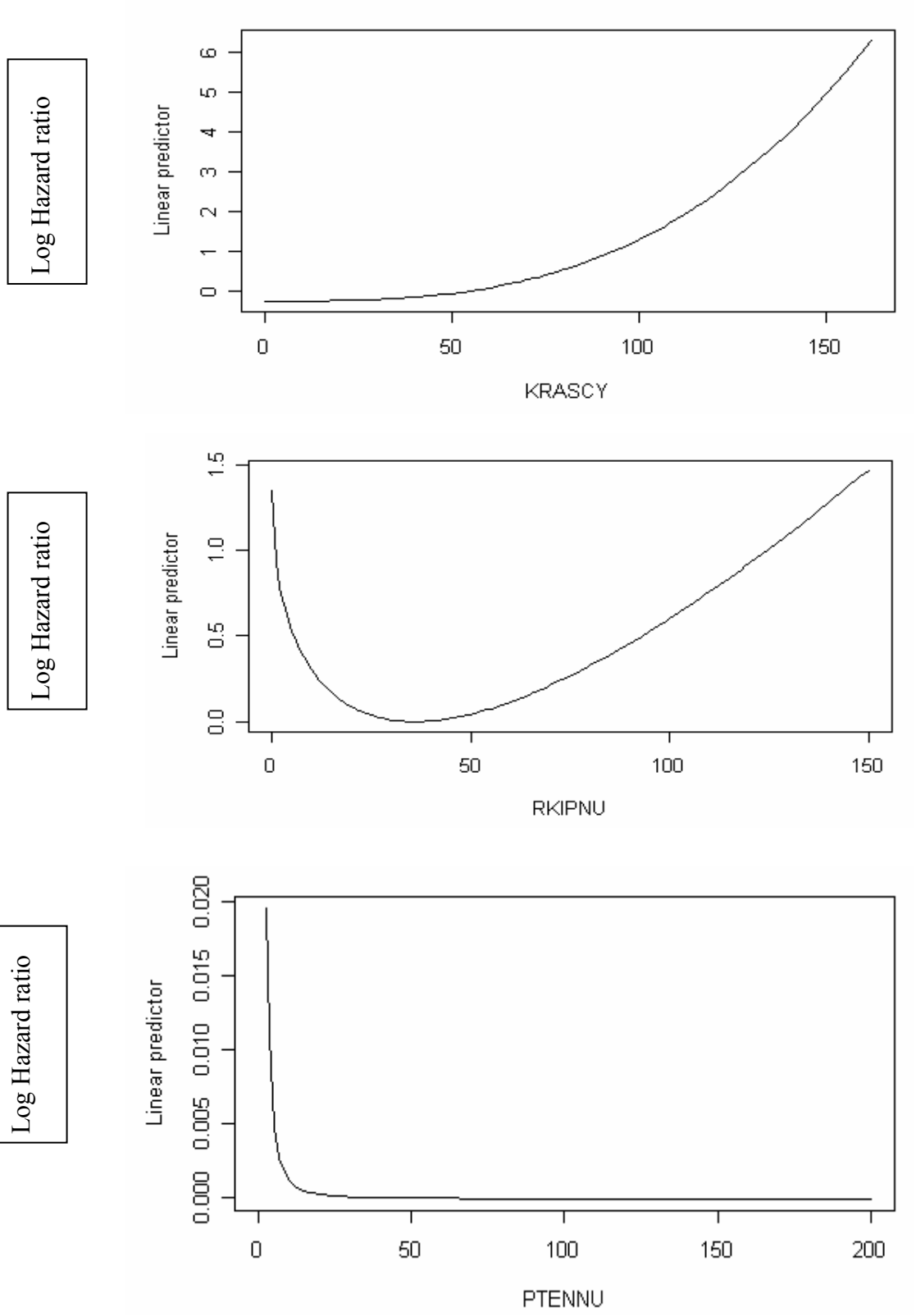
Minimum P-value selected 10 biomarkers as potentially informative. The number of recurrences/ patients in low/ high risk groups is given in Table 7.2. The only informative biomarker missed by FP but selected by minimum P-value method was Pmapknu. For this biomarker, a threshold effect was found at 104 with unadjusted P-value of 0.003. The P-value for this biomarker in FP method was 0.92.

Table 7.2: Univariate P-values and Hazard Ratios (HR)[4] for biomarkers which are selected as informative

Variable	**FP**		**Linear Cox**		**Minimum P-value method**				
	P-value	HR (95% C.I.)	P-value	HR (95% C.I.)	Optimal split	P-value	# events/ patients in risk groups		HR (95% C.I.)
Praf338cy	0.01	2.17 (1.19, 3.94)	0.01	2.17 (1.19, 3.94)	192	0.001	69/ 300	29/ 57	2.72 (1.76, 4.2)
Praf338nu	0.002	2.56 (1.40, 4.68)	0.002	2.56 (1.40, 4.68)	123	0.001	23/ 125	75/ 232	2.11 (1.32, 3.37)
Mapkcy	0.01	1.56 (1.09, 2.22)	0.01	1.56 (1.09, 2.22)	128	0.003	52/ 239	51/ 137	1.78 (1.21, 2.62)
Prhisto	0.007	0.72 (0.56, 0.91)	0.007	0.72 (0.56, 0.91)	20	0.001	67/ 179	43/ 208	0.48 (0.33, 0.70)
Akt2cy	0.06	0.73 (0.52, 1.02)	0.06	0.73 (0.52, 1.02)	190	0.005	96/ 306	12/ 81	0.44 (0.24, 0.80)
Pmtor	0.02	0.59 (0.38, 0.93)	0.02	0.59 (0.38, 0.93)	100	0.001	96/ 318	10/ 72	0.32 (0.17, 0.62)
Tunel	0.07	1.21 (0.99, 1.48)	0.07	1.21 (0.99, 1.48)	105	0.003	73/ 300	31/ 62	1.90 (1.24, 2.91)
Pher2nu	0.07	1.90 (0.95, 3.78)	0.07	1.90 (0.95, 3.78)	80	0.005	85/ 336	19/ 40	2.01 (1.22, 3.31)
Mtor	0.06	1.53 (0.98, 2.39)	0.06	1.53 (0.98, 2.39)	127	0.01	80/ 327	24/ 52	1.78 (1.13, 2.82)
Krascy	<0.001	3.43 (1.94, 6.08)	0.59	0.86 (0.49, 1.49)	7	0.01	21/ 45	89/ 342	0.55(0.34, 0.88)
Rkipnu	<0.001	0.66 (0.52, 0.84)	0.65	1.16 (0.61, 2.24)	8	0.001	30/ 73	79/ 314	0.49 (0.32, 0.75)
Ptennu	0.02	1.13 (1.02, 1.26)	0.83	0.94 (0.51, 1.71)	2.5	0.01	31/ 84	72/ 289	0.61 (0.40, 0.93)
Pmapknu	0.92	1.03 (0.58, 1.83)	0.92	1.03 (0.58, 1.83)	104	0.003	75/ 312	25/ 56	1.98 (1.26, 3.12)
Akt1nu	0.44	0.86 (0.57, 1.27)	0.44	0.86 (0.57, 1.27)	107	0.05	94/ 312	17/ 84	0.60 (0.36, 1.01)

[4] For biomarkers which are selected by FP and Cox methods, but not for Rkipnu and Ptennu, HR shows the amount of increase in risk per 100 unit change in biomarker value
P-value thresholds used for FP, linear Cox, and Minimum P-value were 0.10, 0.10, and 0.005 respectively
For each of 3 methods, grey cells are uninformative biomarkers

Figure 7.2: Shapes of risk function for biomarkers with univariate polynomial effects (Krascy (top panel), Rkipnu (middle panel), and Ptennu (bottom panel)



iv) Application of 'quartile dichotomisation'

Dichotomisation at lower quartile (Q1)

Just 2 biomarkers were informative at Q1 (Praf338nu and Prhisto). These two biomarkers were also informative at Q2 and Q3 (Table 7.3).

Dichotomisation at median (Q2)

Informativeness of 4 biomarkers was revealed applying the split at median, and all of them were also informative at Q3 (Table 7.3).

Dichotomisation at upper quartile (Q3)

Applying the split at Q3, informativeness of 8 biomarkers was revealed (Table 7.3). Only number of patients in the risk groups and HR corresponding to this split is given (Table 7.3).

v) Application of 'non-ordinal quartile dichotomisation'

Having performed 4 comparisons for each variable, and therefore a total of 276 tests, only one biomarker was selected as informative. For Akt1nu, patients with values in the third quartile exhibited different relapse risk when compared with the remaining patients who had Akt1nu values above or below this range.

In the third quartile of Akt1nu, 38 events in 102 patients were observed. In the remainder (those who had Akt1nu lower than median or higher than third quartile) 73 recurrences in 294 patients were seen. The risk of recurrence for patients in the third quartile, in comparison with other patients, was 1.81 (95% C.I.: 1.22, 2.69) with unadjusted P-value of 0.003.

Table 7.3: Quartiles and univariate P-values for biomarkers which are selected as informative and univariate Hazard Ratio (HR) for top quartile (threshold P-value 0.033)

Variable	**Q1**	P-value	**Q2**	P-value	**Q3**	P-value	# events/ patients in risk groups		HR (95% C.I.) for Q3
Praf338cy	136	0.12	167	0.01	190	0.001	59/ 264	39/ 93	2.03 (1.34, 3.05)
Praf338nu	113	0.01	135	0.01	158	0.01	65/ 271	33/ 86	1.78 (1.17, 2.71)
Mapkcy	70	0.17	110	0.11	147.7	0.03	68/ 282	35/ 94	1.55 (1.03, 2.33)
Prhisto	0	0.01	35	0.001	140	0.02	92/ 294	18/ 93	0.56 (0.34, 0.92)
Akt2cy	125	0.37	158	0.03	188	0.01	93/ 294	15/ 93	0.47 (0.27, 0.81)
Pmtor	20	0.40	50	0.40	90	0.02	87/ 301	19/ 89	0.54 (0.33, 0.90)
Tunel	0	0.54	0	0.54	72.5	0.02	65/ 272	39/ 90	1.60 (1.06, 2.40)
Pher2nu	25	0.19	43.3	0.07	65	0.33	75/ 291	29/ 85	1.24 (0.81, 1.91)
Mtor	40	0.41	65	0.17	105	0.024	68/ 287	36/ 92	1.59 (1.06, 2.38)
Krascy	27	0.40	53	0.72	85	0.53	90/ 299	20/ 88	0.86 (0.53, 1.39)
Rkipnu	12	0.16	28	0.65	50	0.59	80/ 293	29/ 94	1.12 (0.73, 1.72)
Ptennu	5	0.06	25	0.80	53.3	0.66	78/ 276	25/ 97	0.90 (0.58, 1.42)
Pmapknu	45	0.11	72	0.99	95	0.15	68/ 279	32/ 89	1.37 (0.89, 2.09)
Akt1nu	30.5	0.57	67.5	0.30	102	0.09	90/ 300	21/ 96	0.67 (0.41, 1.08)

For each of 3 splits, grey cells are uninformative biomarkers

vi) Comparison of selection of biomarkers by different methods

Table 7.4 summarises the selection of biomarkers as potentially informative by different univariate screening methods applied. The FP method was the most inclusive technique and selected the highest number of biomarkers as potentially informative (12 biomarkers), and was the only method captures the association of Rkipnu and Ptennu with RFS (both polynomial). Therefore, selection of biomarkers by other methods was compared to FP. Main finding were as follows:

The linear Cox model selected 9 biomarkers as being informative. Application of this technique resulted in loss of 3 biomarkers with polynomial association.

The minimum P-value method selected a total of 10 biomarkers as being informative. This technique missed to select Mtor (which showed linear) and Krascy and Ptennu (which showed polynomial effect). This was the only method identified effect of Pmapknu

Number of biomarkers significant at Q1 and Q2 were 2 and 4 respectively, all of them were also significant at Q3. Applying the split at Q3, a total of 8 biomarkers were selected. In comparison with FP, 4 biomarkers (Pher2nu and 3 biomarkers with polynomial effects) were not screened in.

'Non-ordinal quartile dichotomisation' suggested only the informativeness of Akt1nu. All other methods missed informativeness of this biomarker.

Table 7.4: Selection of biomarkers, as potentially informative, by different univariate screening methods

Row	Biomarker	FP	Cox	Minimum P-value	Ordinal quartiles (number max 3)	Non-ordinal quartiles (number max 4)	Selection frequency
1	Praf338cy	Yes	Yes	Yes	Yes (2)		5
2	Praf338nu	Yes	Yes	Yes	Yes (3)		6
3	Mapkcy	Yes	Yes	Yes	Yes (1)		4
4	Prhisto	Yes	Yes	Yes	Yes (3)		6
5	Akt2cy	Yes	Yes	Yes	Yes (2)		5
6	Pmtor	Yes	Yes	Yes	Yes (1)		4
7	Tunel	Yes	Yes	Yes	Yes (1)		4
8	Pher2nu	Yes	Yes	Yes			3
9	Mtor	Yes	Yes		Yes (1)		3
10	Krascy	Yes*					1
11	Rkipnu	Yes*		Yes			2
12	Ptennu	Yes*					1
13	Pmapknu			Yes			1
14	Akt1nu					Yes (1)	1
	# of detected informative biomarkers	12	9	10	8	1	

* Polynomial form

7.4.2 Selection of informative biomarkers and form of association

As summarised in Table 7.4, the first 7 biomarkers (rows 1 to 7) were selected by 4 techniques. Following the methods explained in section 7.3.2, I will keep them in the continuous form and adopt the linear risk function.

Two biomarkers (rows 8 and 9) were selected by 3 techniques. These two biomarkers will also be used in continuous form with linear risk function.

Rkipnu (row 11) was selected by FP and minimum P-value methods. I will use form selected by FP. The rest of biomarkers (rows 10, 11, and 14) were selected by only one method and will be included in the selected form.

None of 3 categorical biomarkers were selected as being informative. Table 7.5 lists functional form for each variable which univariately predicted RFS. Number of missing values for each variable and statistics for distribution are also given. The rate of missing value for selected variables varied from 1.2% (Akt1nu) to 11% (Praf338cy & Praf338nu).

Table 7.5: Distributional statistics and rate of missing value for the biomarkers selected as potentially informative

Thesis abbreviation	Form of risk function	Min	Q1	Q2	Q3	Max	Missing (%)
Praf338cy	Linear	58	136	167	190	275	11%
Praf338nu	Linear	5	113	135	158	220	11%
Mapkcy	Linear	0	70	110	147	260	6.2%
Prhisto	Linear	0	0	35	140	300	3.5%
Akt2cy	Linear	0	125	158	188	275	3.5%
Pmtor	Linear	0	20	50	70	90	2.7%
Tunel	Linear	0	0	0	72	400	9.7%
Pher2nu	Linear	0	25	43	65	100	6.2%
Mtor	Linear	0	40	65	105	190	5.5%
Krascy	Polynomial	0	27	53	85	162	3.5%
Rkipnu	Polynomial	0	12	28	50	150	3%
Ptennu	Polynomial	0	5	25	53	200	7%
Pmapknu	Threshold	0	45	72	95	180	8.2%
Akt1nu	Non-ordinal	0	31	67	102	250	1.2%

7.5 Discussion

7.5.1 Comparison of screening methods

The reason I applied a range of statistical methods and explored form of association was that, in this work, my aim was to supply information to a cancer biologist. This might help him to generate new hypotheses about the role of biological variables on the course of breast cancer disease and to enhance the understanding of aetiology of breast cancer.

In total 14 biomarkers are selected as being informative. Applying 5 screening methods, FP was the most inclusive approach. This approach missed selecting only Pmapknu (with a threshold effect) and Akt1nu (with a non-ordinal effect).
Since FP explores a wide range of power transformations, it might be over-inclusive by finding artificial and/ or unstable associations. Furthermore, form a biological perspective, results must be interpretable (see below).

Another interesting finding was that when I dichotomised biomarkers at bottom, middle, and top quartiles, I saw that the quartile at which biomarkers were dichotomised played an important role in informativeness of biomarkers. Upper quartile was the most inclusive method, possibly due to positively skewed distribution of biomarkers.

7.5.2 Biological interpretability of detected non-linear forms

For modelling, 3 biomarkers were selected with polynomial risk function (Krascy, Rkipnu, and Ptennu).

The association between Krascy and RFS was best captured by cubic transformation. When the association between Rkipnu and RFS was discussed with the biological collaborators, they explained that the shape of association might be interpreted in two ways: a surrogate effect (left-hand side of Figure 7.2) and also the specific effect of the covariate (right-hand side of Figure 7.2). By surrogate effect they meant an indicator of the effect of another variable also important to RFS. Given that this screening step was univariate, the information of this variable might disappear in the multifactorial analysis, when other biomarkers are included in the model.

I explained the biological collaborators that the corresponding graph for Ptennu suggested a threshold effect at very low histoscore level. Professor John Bartlett confirmed that it is plausible. Furthermore, from biological perspective, threshold effect of Pmapknu and non-ordinal effect of Akt1nu are plausible. It is plausible that after a threshold value, biomarker predicts the outcome. It is also biologically possible that patients in one of quartiles, in comparison with the remainder, show different survival curve and therefore predict outcome.

7.5.3 Informative biomarkers

One important aim of screening procedures applied was to select a reduced number of informative biomarkers to offer to a subsequent multifactorial model (see Chapters

8 and 10 for details). However, I was also concerned not to miss biomarkers that might be important and therefore a range of screening methods were applied to select biomarkers and form together.

I was aware that the FP can fairly accommodate threshold and even complex v-shape associations. The reason why the biological collaborators of this research asked me to check for complex non-ordinal associations, was that Tovey *et al.* showed that 'the frequency distribution for the levels of HER2 expression demonstrates the bimodal expression pattern, with a nadir at ten times the normal expression level' [Tovey SM et al., 2006]. Patients with 1-10 times normal tissue had the best survival while those with low normal tissue and high values (> 10 times normal) together had the worst survival [Tovey SM et al., 2006]. This suggests that it is possible to find a subgroup of patients in the middle of the distribution with different survival experience. In the data set I analysed, HER2 was not associated with RFS. One explanation might be the fact that this cohort comprised only ER+ patients.

No one of the 5 techniques applied was able to select all 14 univariately informative biomarkers. Therefore, application of all techniques might be required when the aim is to select potentially informative biomarkers which had complex associations with outcome, to describe the sample as best as possible, and to understand more about biology of disease.

At this stage, multiple testing (due to multiple comparisons undertaken) is not an issue since I wanted to be inclusive and give variables every chance to unscreened variables to be selected for modelling. However, in development of models

reproducibility of non-linear effects detected will be checked and unstable associations will be excluded from analyses (see Chapter 8 for details). I will develop multifactorial model with and without use of biomarkers missed by FP, to address their contribution to the performance of the models.

7.5.4 Selection of informative biomarkers and form for multifactorial modelling

As explained in section 7.3.2, one of the aims of screening procedures applied was to select a reduced set of biomarkers for modelling (see Tables 7.2 and 7.3). However, majority of biomarkers were selected by more than 1 screening technique (see Table 7.4). Therefore, a decision abut form is also required. In models developed in Chapter 8, I will focus on techniques which optimise the form. This is because my aim is to describe the sample as best as possible. However, models developed will be challenged in Chapter 10 to investigate whether selection of simpler forms was adequate.

7.6 Chapter summary

- Fractional Polynomial was the most inclusive technique in terms of selection of informative biomarkers. Only 1 threshold and 1 non-ordinal association were missed by FP.

- Non-FP methods might extract information on risk curves more complicated than those available within FP family. Application of such methods might generate new questions about biology of disease.

- Multi-faceted screening process is required to ensure that all potentially informative variables are selected for modelling.

Chapter 8 PROGNOSTIC MODELLING OF MANY SKEWED VARIABLES WITH MISSING DATA

8.1 Introduction

As noted in Chapter 4 (section 4.1), the main 4 practical challenges in the development of prognostic models are to deal with many variables, to detect appropriate form of association, to impute missing data, and to assess internal validity of model guarding against instability and overfitting.

The aim of this chapter is to fit the best possible multifactorial models, by combining suitable data reduction techniques, risk function detection methods, and missing value imputation approaches, followed by a stability checking procedure to tackle overfitting.

8.2 Aim

The main aim of this part of the research is to develop pragmatic strategies to be used in modelling when large number of skewed variables with missing data is available.

The main objectives are to:

1. Develop a multifactorial regression model without use of biological knowledge

2. Make the use of biological knowledge in model development of the regression model and to address its role on model performance

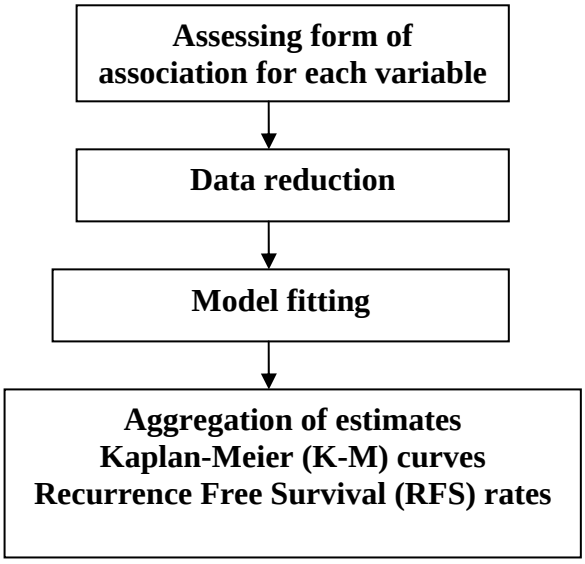
3. Construct a graphical decision tree

8.3 Methods

In development of the multifactorial models presented in this chapter, I took two overall approaches: regression modelling (without and with incorporation of biological knowledge) and Tree-based Survival Methods (TSM).

Two regression-based models were developed. The overall process of model development is shown in Figure 8.1. The main difference between 2 approaches was the mechanism applied to restrict number of variables being offered to the multifactorial model. I applied 2 selection strategies: a statistically approach which I called Univariately Informative Variable Selection (UIVS) and Biologically Guided Variable Selection. These will be now in more details (Figures 8.2 and 8.3).

Figure 8.1: Overall approach in development of two regression models



8.3.1 Univariately Informative Variable Selection (UIVS) Model

This is an extension of the conventional screening method, involving also detection of appropriate form of association, preceded by imputation of missing data, and followed by bootstrapping (Figure 8.2).

i) Assessment of form

Details of detection of form of association are explained in section 7.3.1.

ii) Data reduction

Details of selection of univariately informative biomarkers and appropriate form are explained in sections 7.3.2. Biomarkers and forms which are listed in Table 7.5 were candidate for the UIVS Model. Clinical variables (tumour size, grade (1, 2, 3), and nodal stage (1, 2, 3)) were also submitted to the model. Although tumour size had a continuous distribution to attempt was made to optimise form of association. Furthermore, this variable was not dichotomised. This is because in original development of NPI, no transformation was applied to tumour size and I wanted to investigate whether biomarkers would retain in the multifactorial models in presence of clinical variables used in NPI.

iii) Model fitting

Data imputation

Set of reduced biomarkers, clinical variables, and RFS outcome were submitted to the imputation model. Applying the MICE method [Van Buuren S et al., 1999], missing were imputed 10 times (section 4.4.4). Transformations derived in screening phase (Chapter 7) were applied to all 10 imputed data sets.

Bootstrap samples

To circumvent the risk of an over-fitted model, I have employed bootstrap sampling to refine the models by excluding variables with unstable form, or unreliably included as necessary for prediction. A total of 100 bootstrap samples were drawn from each of 10 imputed data sets, leading to 1000 data sets in total.

Elimination of unstable predictors with threshold or non-ordinal effects

Since there was no multifactorial procedure to the link 3 screening methods applied in 'assessment of form' step (FP for linear or polynomial effects (section 4.4.2), minimum P-value for threshold effects (section 4.4.3), and non-ordinal dichotomisation for non-ordinal effects), stability of threshold effects and non-ordinal effects were assessed across 1000 samples univariately. Biomarkers were dropped if form was replicated in < 50% of samples.

Elimination of predictors with unstable polynomial effects

Optimum form for biomarkers with univariate polynomial association was ascertained applying Multivariate Fractional Polynomial (MFP) [Royston P and

Sauerbrei W, 2008] to 1000 samples. A variable with univariate polynomial association might show different forms of association in bootstrap samples. If optimum risk function (linear, FP1, FP2) was replicated in < 50% of samples, the biomarker was dropped from further consideration.

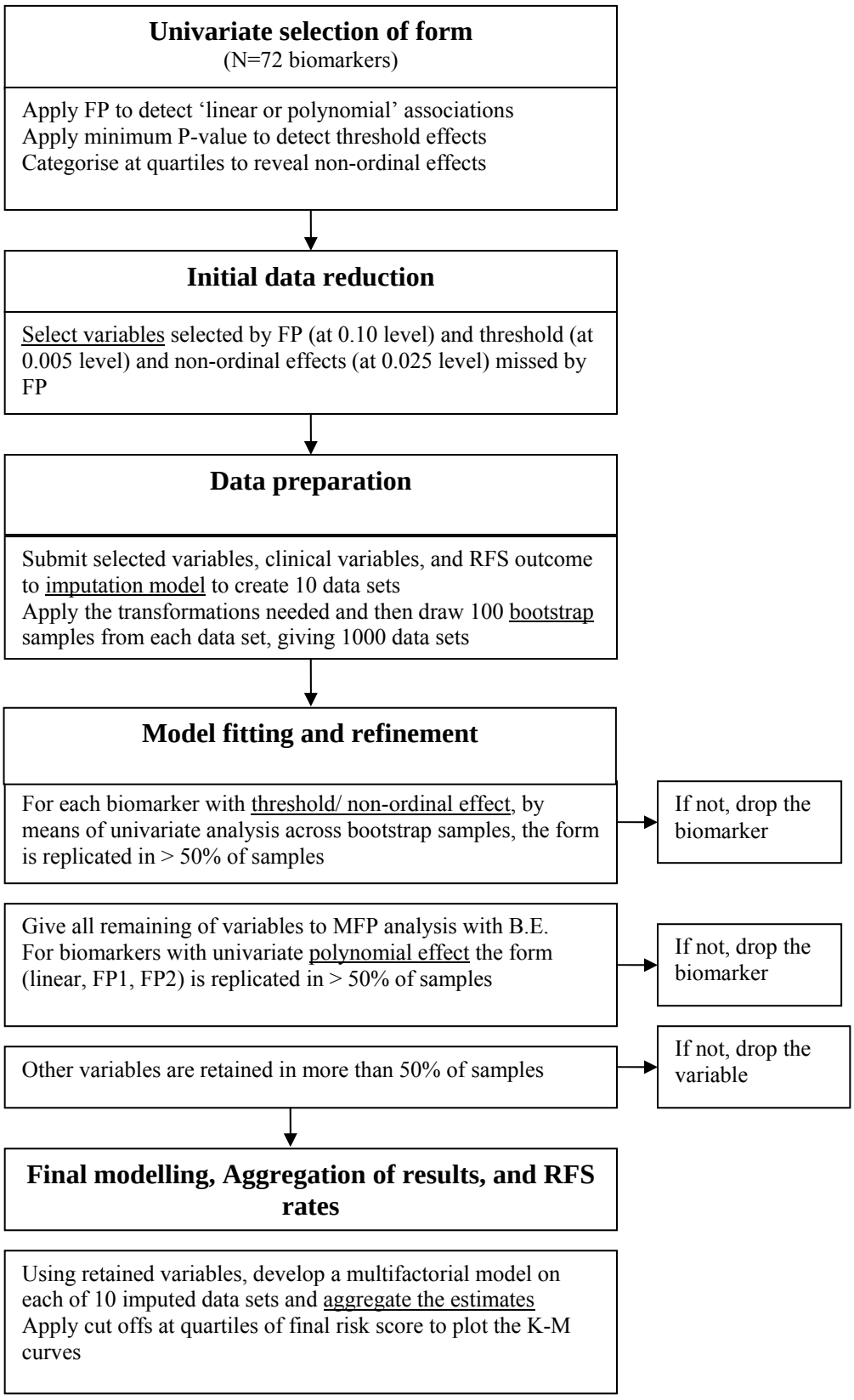
Elimination of unreliable predictors with low inclusion frequency

Besides predictors with unreliable polynomial association; clinical variables, transformed version of biomarkers with stable threshold or non-ordinal effects, and biomarkers with univariate linear association were removed if retained in < 50% of samples.

iv) Aggregation of results

For variables retained, estimated regression coefficients and Standard Errors (S.E.'s) were aggregated across 10 imputed data sets (as described in methods section 4.4.4 part i). In the case power(s) selected for variables with stable polynomial effect were not the same across 1000 samples, the most frequent power seen were applied. A final risk score was calculated (section 4.4.4 part iv) and patients were categorised into 4 risk groups (section 4.4.5).

Figure 8.2: Process of development of the UIVS Model



8.3.2 Biologically Guided Variable Selection (BGVS) model

In the Biologically Guided Variable Selection (BGVS) approach, modelling was performed within family sets created by Professor John Bartlett. The process continued as follows (Figure 8.3).

i) Families

Seven families were formed on the basis of presumed pathway to tumor progression. Remainder of biomarkers comprised the eighth set ('Non-family' biomarker set).

ii) Assessment of form

Regarding the form of association, there was scant biological knowledge as the appropriate form. Therefore, preliminary used form selection in screening phase (Chapter 7) and used in the UIVS Model were taken.

iii) Dimension reduction

Since family sets were specified by Professor John Bartlett, instead of working with 72 biomarkers, I worked on substantive sets of biomarkers, each with reasonably smaller number of biomarkers. For each biomarker set I was therefore able to develop a multifactorial model. I then established for each biomarker family a combination of the family variables that constituted an informative and parsimonious index (risk score) to predict Recurrence Free Survival (RFS), as explained below.

iv) Development of multifactorial models and estimation of family risk scores

For each of 7 biomarker family sets, missing data were imputed 10 times [Van Buuren S et al., 1999] (section 4.4.4). Then, within each family set, multifactorial models in conjunction with B.E. variable selection method were developed as explained below.

1) For family sets in which none of biomarkers predicted RFS univariately or only univariate linear forms detected, multifactorial Cox model was used (BAD, PgR, and HER families).

2) For family sets in which univariate polynomial and/ or linear effects were detected, MFP was applied (RAS, MTOR families) [Royston P and Sauerbrei W, 2008]. In the case optimum form was replicated in > 5 samples but power(s) was not consistent across data sets, I applied the power replicated in the majority of samples.

3) In family sets with either of threshold or non-ordinal effects, and linear effects detected, multifactorial Cox model was applied (AKT and MAPK families).

Estimates were aggregated as explained in section 4.4.4 parts i and ii. Stability of the form and inclusion frequency was checked across 10 imputed data sets and only those replicated in > 5 samples (50%) were applied. Risk scores were derived as explained in section 4.4.4 part iv.

v) Model fitting

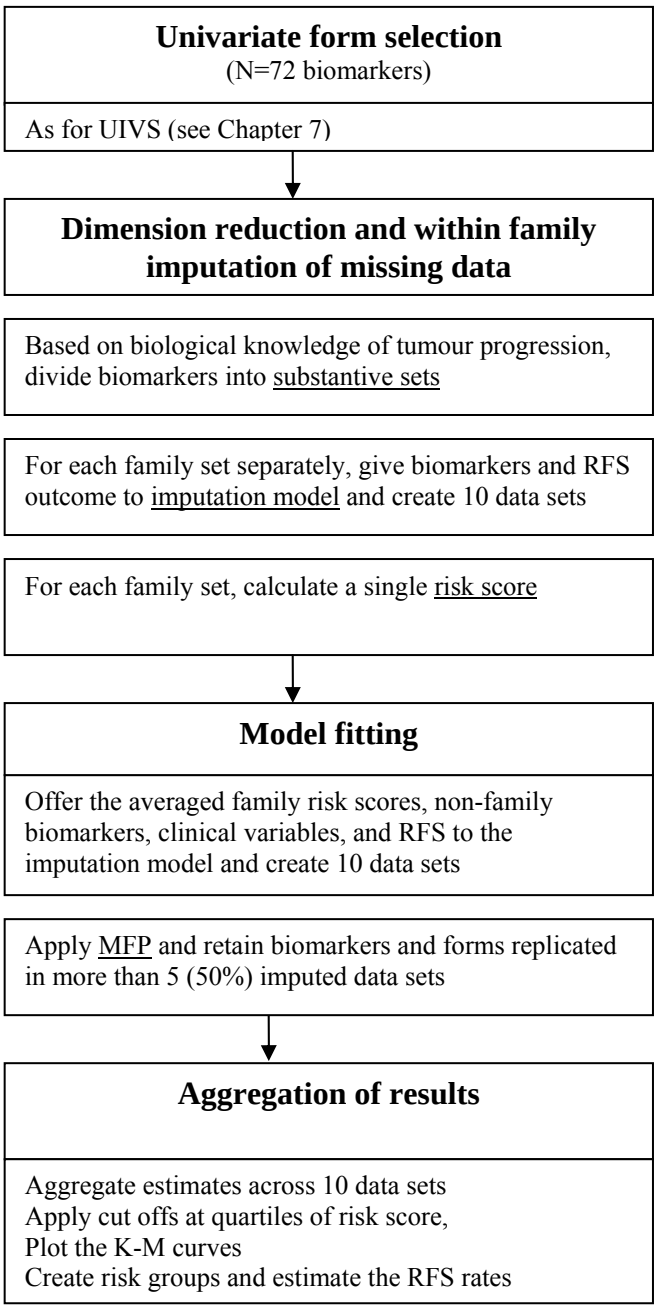
In the second stage the averaged risk scores were offered to the model, plus the clinical predictors (grade, tumour size and nodes) and non-family TMA variables. However, missing data for non-family TMA biomarkers and clinical variables were not imputed in the first round which was restricted to the family.

At the second stage, again, 10 data sets were imputed, for the non-family and clinical variables [Van Buuren S et al., 1999] (section 4.4.4). The family indices being offered were already averaged across data sets, so in each of the modelling runs their values were constant across the 10 data sets. Since one of the single variables (Rkipnu) had univariate polynomial association with outcome, MFP was applied [Royston P and Sauerbrei W, 2008].

vi) Aggregation of results

Risk scores, clinical variables, and non-family biomarker sets were candidate for the multifactorial model. A multifactorial model was fitted to each of 10 imputed data sets and results were aggregated as explained in 4.4.4 part i. Variables retained in more than 5 samples were selected for the final model.

Figure 8.3: Process of development of the BGVS Model



8.3.3 Tree-based Survival Method (TSM)

The approaches proposed so far require preliminary steps to impute the missing data and to reduce the number of variables. This makes the process of model building very complex and time consuming. A far easier approach would be to construct a decision tree. The main steps for this model were follows.

i) Form of association and dimension reduction

In this approach variables are routinely transformed into binary versions and so no decision was required on the issue of form.

Furthermore, in TSM there is no limitation about number of variables so no preliminary step was needed to tackle issue of many candidate variables.

ii) Construction of decision tree and dealing with missing data

TSM is an extension to minimum P-value method, which was explained in 4.4.3. Log-Rank test was applied to every possible cut point for each prognostic variable (but not the outer 20% values in distributions) so as to select the split with the most difference in outcome between two groups [Williams BA et al., 2006].

To allocate patients with missing data into the appropriate group, the 'surrogate variable' approach was used by re-applying the partitioning algorithm. If the surrogate variable has the missing value on the same subject, then a second surrogate variable was used and so on [Therneau TM and Atkinson EJ, 1997].

The whole process was continued creating a tree structure until final subgroups (terminal nodes) with a minimum size of at least 30 patients were achieved.

iii) Refinement of tree

To tackle potential overfitting, due to multiple testing undertaken, branches with P-value higher than 0.002 (corresponding to 0.05 if one single test applies) [Altman DG et al., 1994] was omitted [Radespiel-Troger M et al., 2003; Dannegger F, 2000].

iv) Amalgamation of groups with similar survival curve

Although TSM ensures that the two terminal nodes within a branch are significantly different, it remains possible that terminal subgroups from distinct branches might have very similar survival curves. In addition, number of patients in some of the final nodes might be low leading to un-robust estimation of event free rates. Therefore, further examination is required of survival curves and event-free rates of terminal nodes, followed by amalgamation of subgroups with similar curves [Segal MR and Bloch, 1989; Banerjee M et al., 2004].

8.3.4 Comparison of approaches

Models developed were compared as explained in section 4.4.7. In addition, concordance between the UIVS and BGVS indices was checked by plotting of Bland-Altman graph.

If a new prognostic model were to be used in clinical management of breast cancer, then it is likely that initially at least it would be used in parallel to Nottingham Prognostic Index (NPI), which is the established gold standard prognostic index in UK. Therefore when NPI designated a patient as ‘high risk’, management would be unlikely to neglect this, even if the new model indicated the patient was not high risk.

Therefore, initially the models would be used only for risk stratification of those patients who were not deemed high risk by NPI.

To ascertain performance in this regard, I excluded the patients who are classified as high risk based on standard NPI (NPI>5.4). New quartiles for the UIVS and BGVS risk score were calculated and applied as cut points to investigate whether it was still possible to stratify patients into well diverged risk groups.

I also checked ability of the risk groups derived from the UIVS and BGVS indices to predict other end points, in particular Recurrence Free while on Tamoxifen treatment (RFoT) and Overall Survival (OS). K-M survival curves and event free rates are given in Appendix 1.

8.4 Results

8.4.1 The Univariately Informative Variable Selection (UIVS) Model

In total 14 biomarkers (9 linear, 3 polynomial, 1 threshold, and 1 non-ordinal) and 3 clinical variables were candidates for the UISV Model. Investigation of the threshold effect of Pmapknu found that in about 55% of bootstrap samples the optimal split was around 104. Additionally, the median and mode of selected optimal thresholds was 104. Investigation of the non-ordinal effect of Akt1nu found that in 80% of replications, the patients in the third quartile of Akt1nu significantly differed with the remainder. Therefore these two biomarkers were selected for the multifactorial model in the transformed version, and both were retained in > 50% of the samples (Table 8.1).

When MFP was applied, 2 biomarkers (Rkipnu and Ptennu) had unstable forms of risk function (the optimum form was replicated in < 50% of samples), and were therefore excluded (Table 8.1). There were 7 other clinical variables with inclusion frequencies of <50%, which were also excluded (Table 8.1). Prhisto was retained in about 45% of replications, suggesting borderline effect of this variable. Grade was retained in the model in only 20% of replications.

The final model was therefore developed using information on 6 biomarkers and 2 clinical variables. For biomarkers the Hazard Ratios (HR) shows the amount of increase in risk of recurrence per 100 unit change in the biomarker histoscore.

Table 8.1: Estimated hazard ratios and inclusion frequency of variables in the UIVS Model

Variable (Family)	HR (95% C.I.)	P-value	Inclusion frequency
Nodal stage (clinical)	1.82 (1.38, 2.40)	<0.001	98.0%
Size (cm) (clinical)	1.21 (1.10, 1.31)	0.001	95.2%
Pmtor (MTOR)	0.33 (0.19, 0.59)	<0.001	79.1%
Tunel (Non-family)	1.49 (1.23, 1.81)	<0.001	85.1%
Praf338cy (MAPK)	2.12 (1.07, 4.02)	0.03	70.8%
Pmapknu (MAPK)	2.8 (1.72, 4.57)	<0.001	79.0%
Krascy (RAS)	6.05 (2.23, 16.44)	<0.001	71.5% FP2 8.1% FP1 2.0% Linear
Akt1nu (AKT)	0.54 (0.36, 0.82)	<0.001	92.3%
Prhisto (PgR)	a.		44.0%
Praf338nu (MAPK)	a.		15.6%
Grade (clinical)	a.		22.0%
Mtor (MTOR)	a.		18.4%
Mapkcy (MAPK)	a.		8.6%
Akt2cy (AKT)	a.		10.5%
Pher2nu (PgR)	a.		10.0%
Rkipnu (Non-family)	b.		30.4% FP2 13.6% FP1 11.6% Linear
Ptennu (MTOR)	b.		27.2% FP2 19.9% FP1 13.4% Linear
Performance			
C-index	79%		
R-square	28%		
Chi-square	126.5		
NR Index in comparison with NPI^{q4}	18.3% (0.02)		

a. Excluded as inclusion frequency was <50%
b. Excluded as inclusion frequency of form was <50%
NPI^{q4}: Standard NPI risks score categorised into 4 risk groups by applying quartiles as cut offs

8.4.2 The Biologically Guided Variable Selection (BGVS) Model

Out of 7 families, 2 did not find any variables to be statistically significant in predicting recurrence (BAD, HER). In the PgR and RAS families, only a single biomarker predicted the outcome, these were: Prhisto with linear form in PgR, and Krascy with FP2 in RAS. These 2 biomarkers were retained in all 10 imputed data sets. The FP2 variable captured the effect of Krascy in all 10 data sets where optimum powers were (3, 3) (Table 8.2).

In the MTOR family, both Mtor and Pmtor were retained in the multifactorial model and contributed to the "mTOR" biomarker set risk score. Ptennu showed a univariately polynomial effect, but lost this polynomial effect in the multifactorial model and was not used in the family's risk score.

In the AKT family, the non-ordinal effect of Akt1nu was stable and replicated in all 10 imputed data sets, and therefore a non-ordinal transformation was applied. Akt2cy and Akt1nu were retained in the multifactorial model and contributed to the derived risk score (Table 8.2).

In the MAPK family, the threshold effect of Pmapknu was the same in all 10 imputed data sets, and therefore the binary version of this biomarker was used. Praf338cy and binary version of Pmapknu were elements of the risk score for the MAPK biomarker set.

The BGVS Model was derived using: the final risk scores, the single biomarkers identified as potentially informative (Prhisto, Krascy), the clinical variables, and the TMA biomarkers that were not included in any set. This resulted in AKT, MTOR, and MAPK biomarker indices, as well as Tunel, Krascy, nodal status, and tumour size being retained in the final model.

Table 8.2: Variables which contributed to the estimate of the risk of recurrence, separately by biomarker family, and corresponding hazard ratios

Family	Estimated risk score/ individual variables	HR (95% C.I.)
BAD	----------	----------
HER	----------	----------
PgR	Prhisto	----------
RAS	-0.006 x {(Krascy/10) ^ 3} -0.002 x {(Krascy/10) ^ 3} x { Ln (Krascy+10) /10}	2.74 (1.71, 4.38)
AKT	-0.64 x Akt1nu* - 0.003 x Akt2cy	1.95 (1.07, 3.59)
MAPK	0.62 x Pmapknu* + 0.007 x Praf338cy	3.01 (1.71, 5.32)
MTOR	0.006 x Mtor - 0.006 x Pmtor	3 (1.88, 4.81)
Non-family	Tunel**	1.35 (1.09, 1.67)
Clinical	Nodal stage Tumour size (cm)	1.92 (1.46, 2.52) 1.14 (1.10, 1.20)

* binary form
** For tunel, HR per 100 unit change is given

8.4.3 Tree-based Survival Model (TSM)

The constructed tree is shown in Figure 8.4, with ovals showing the terminal nodes. The numbers in each node represent the number of recurrences and patients in that node. For each split, the P-value corresponding to that of the Log-Rank test is given.

A total of 5 variables were used to construct the tree. The first two variables, which best separated the patients, were nodal status and tumour size. The three biological variables Tunel, Prhisto, and Krascy were also required.

The absolute difference between estimated 7 RFS rates, those of nodes 3 and 4, was 20% (Table 8.3). When comparing nodes 1 with 2, and 4 with 6, the corresponding rates were much lower (6% for each comparisons). Furthermore, the number of patients that formed nodes 1 and 2, and nodes 5 and 6, was less than those that formed nodes 3 and 4. Therefore, to have more robust estimates, nodes 1 and 2 were combined to create the lowest risk group; and nodes 5 and 6 were grouped to create the highest risk groups.

Figure 8.4: Classification tree using biomarkers and clinical predictors

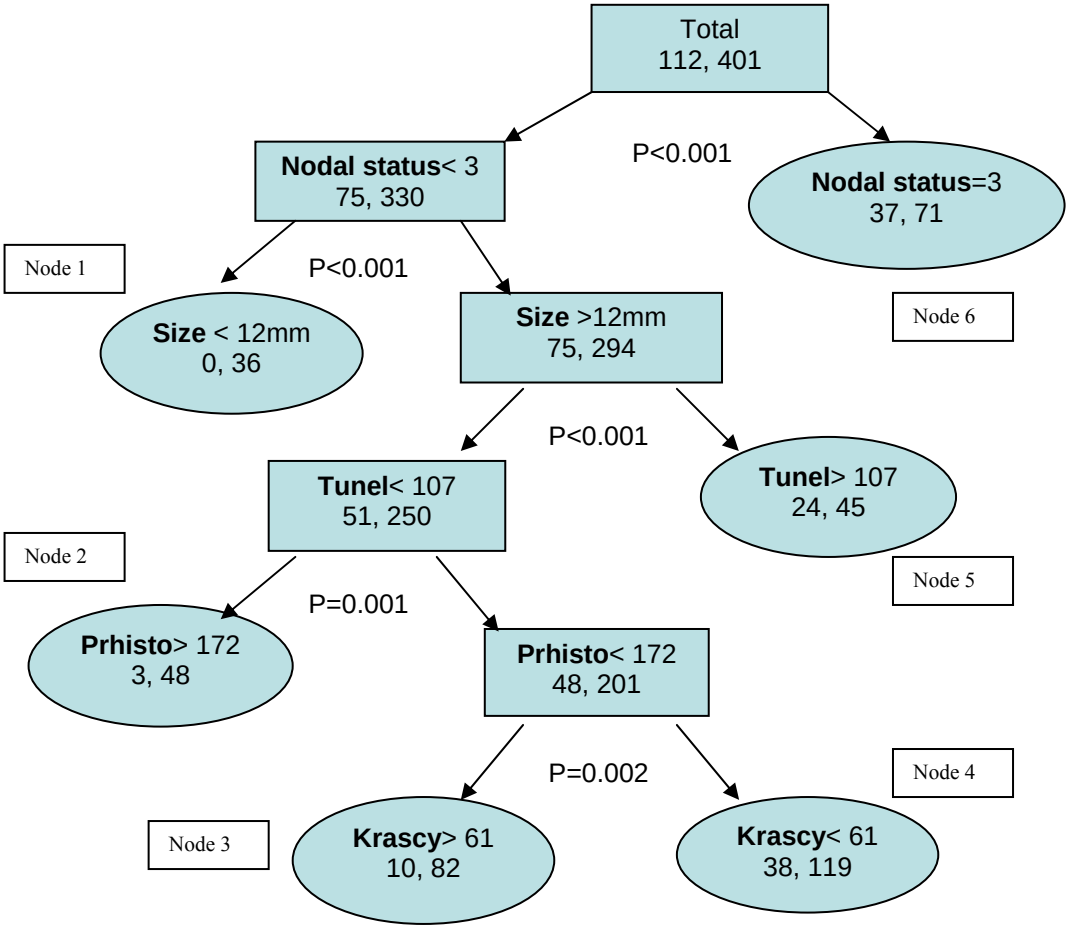


Table 8.3: Estimated RFS rates in each of tree nodes

Node number	N at stat	5-year RFS	7-year RFS
1	36	100%	**100%**
2	48	94%	**94%**
3	82	93%	**87%**
4	119	74%	**67%**
5	45	64%	**54%**
6	71	48%	**48%**

8.4.4 Comparison of the multifactorial models with NPI

Estimated RFS rates in the lowest and highest risk groups are summarised in Table 8.4. Statistics for all models developed are summarised in Table 8.5. K-M curves indicating the risk stratification ability of all the developed models is plotted in Figure 8.5.

i) Comparison between the UIVS with NPI Models

The UIVS Model contained 8 variables. The 7-year RFS rate in the lowest risk group of NPI^{q4} was 91%. Only 4 recurrences were observed among patients in the lowest quartile of the UIVS index, giving a 7-year RFS of 95% (95% CI: 89%, 100%) (Table 8.4). All observed recurrences happened within the first 7 years of follow-up, and therefore the RFS rate remained constant up to the tenth year.

Furthermore, a noticeable improvement in PSEP was seen (55% for UIVS risk grouping versus 42% for NPI^{q4}), indicating that in the UIVS model the lowest and highest risk groups were better distinguished than in the NPI^{q4} model (Table 8.4).

As expected, discrimination, predictive ability, and goodness of fit of the UIVS index was higher than NPI (C-index: 79% versus 72%; R-square: 28% versus 14%; Chi-square 126.5 versus 59.8) (Table 8.5).

The UIVS risk groups, relative to NPI which had 4 equal size risk groups (NPI^{q4}), shifted 16 of 112 recurred patients (30%) into a higher risk group, and 12 non-recurred patients (16%) into a lower risk group (see Appendix 2). This gave a net

gain of 14 percentage points. Among 289 non-recurred patients, 31% were moved into a more appropriate risk group, and 27% were moved into less appropriate risk group, giving a net gain of 4 percentage points. The NR Index (section 4.4.7 part ii) was estimated at 18% (P-value=0.02) (Table 8.5).

Finally I applied the cut offs of the UIVS risk score to create risk groups containing 133, 199, and 69 patients representing low, intermediate, and high risk groups. This is the same as standard NPI with 3 unequal risk groups (NPI^{std3}). Estimated 7-year RFS in the lowest risk group remained at 95% (95% C.I.: 91%, 99%), the corresponding figure at 10-years was 92% (95% C.I.: 86%, 98%). This indicated that 33% of patients had sufficiently low risk of recurrence of breast cancer at 7 years.

ii) Comparison between the BGVS and NPI Models

In total, 10 variables (8 biomarkers and 2 clinical variables) contributed to the BGVS model. A very specific low risk group was detected with only 3 recurrences, all of which happened within the first 3 years of follow-up. Among patients in the lowest risk quartile of the BGVS index, the estimated 7-year RFS rate was 98% (95% CI.: 96%- 100%). This was 7 percentage points higher than that of NPI^{q4} (Table 8.4), and grew to 19% at 10-years. Additionally, a greater than 15% improvement in PSEP was seen (58% for BGVS versus 42% for NPI^{q4}) (Table 8.4).

The discrimination ability of BGVS index was noticeably higher than NPI (79% versus 72%) (Table 8.5), this was also found for predictive ability (27% versus 14%) and goodness of fit (120.3 versus 59.8).

The proportion of recurred patients that moved into more appropriate and less appropriate risk groups were 33% and 12%, respectively (see Appendix 2). The corresponding figures for non-recurred patients were 36% and 25% (Table 8.5). The net gain in the proportion of patients that were reclassified was 21% for those that recurred and 11% for those that did not recur. The NR Index was estimated at 32 % (P-value= 0.001).

When cut offs were applied at the BGVS risk score to create risk groups similar to NPI^{std3}, the estimated 7 and 10-year RFS in the lowest risk group was 96% (95% CI: 92%, 100%) suggesting that 33% of subjects had sufficiently low risk of disease recurrence at 7 years.

iii) Comparison of the TSM and NPI Models

In the lowest risk group identified by TSM, only 3 recurrences out of 84 patients occurred. This gave a 7-year RFS of 96% (95% C.I.: 92%, 100%). The separation ability of TSM and NPI was the same (C-index 72%), however the predictive ability of TSM was marginally higher (R-square 16% for TSM versus 14% for NPI).

The separation ability between the groups depends to a considerable extent on the size of the risk groups. In TSM analysis, the numbers of patients in each of the 4 risk groups were 85, 82, 119, and 115 (lowest to highest risk), whereas NPI had 4 equal size risk groups. Therefore no further comparison between models was made.

Table 8.4: Comparison of estimated RFS rates in the lowest and highest risk groups of models developed and NPI

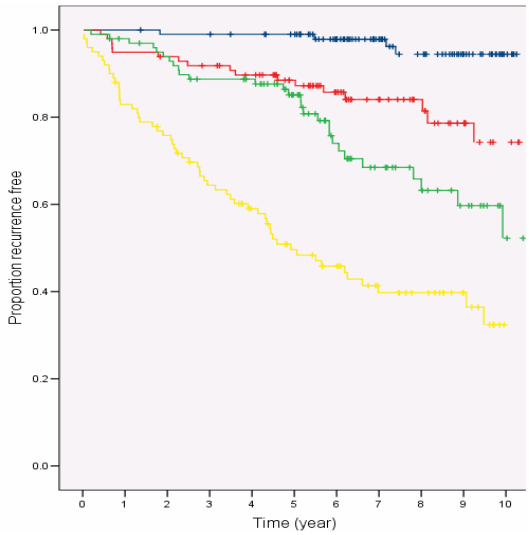
Risk group	Index	N at stat	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Low	NPI	102	95% (91%, 99%)	**91% (85%, 97%)**	84% (72%, 96%)
	UIVS	103	98% (97%, 100%)	**95% (89%, 100%)**	**95% (89%, 100%)**
	BGVS	101	98% (96%, 100%)	**98% (96%, 100%)**	98% (96%, 100%)
	TSM	85	96% (92%, 100%)	**96% (92%, 100%)**	96% (92%, 100%)
High	NPI	100	54% (44%, 64%)	**49% (39%, 59%)**	41% (27%, 55%)
	UIVS	100	46% (36%, 56%)	**40% (30%, 50%)**	31% (17%, 45%)
	BGVS	100	45% (35%, 55%)	**40% (30%, 50%)**	27% (15%, 39%)
	TSM	115	53% (43%, 63%)	**50% (40%, 60%)**	45% (35%, 55%)
PSEP for NPI[q4]			41%	**42%**	43%
PSEP for UIVS risk groups			52%	**55%**	64%
PSEP for BGVS risk groups			53%	**58%**	71%
PSEP for TSM risk groups			43%	**46%**	51%

Table 8.5: Comparison of performance of approaches applied to stratify patients into risk groups

Approach	C-index	R-square	Chi-square	≠recurrences in the lowest-risk group	Net Reclassification Improvement (NR Index) relative to NPI								
					Recurred patients			Non-recurred patients			NR Index		
					Number (%)	Z	P-value	Number (%)	Z	P-value	%	Z	P-value
UIVS	79.0%	28%	123.6	4	16 (14%)	2.23	0.02	12 (4%)	0.90	0.36	18%	2.37	0.02
BGVS	79.0%	27%	120.3	3	23 (21%)	3.42	0.006	30 (11%)	2.26	0.02	32%	4.01	0.001
NPI	72.0%	14%	59.8	10									

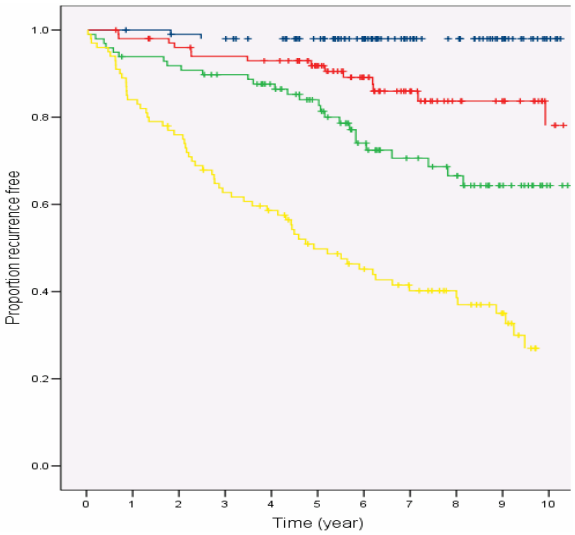
Figure 8.5: K-M survival curves for the UIVS (top left), BGVS (top right), TSM (bottom left), and NPI risk groups (bottom right)

UIVS



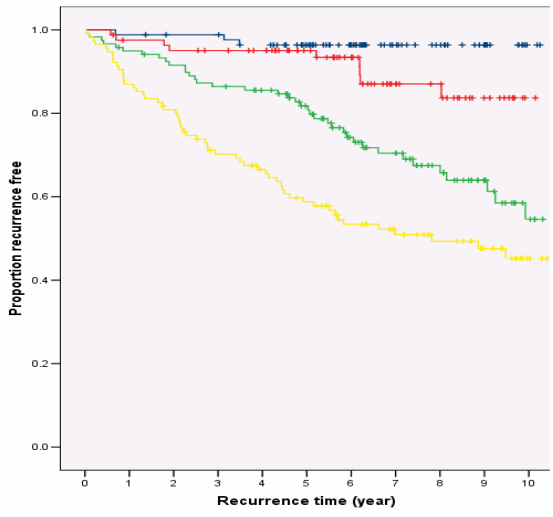
103	101	96	63	37	21
98	88	70	32	23	14
100	84	63	32	16	7
100	61	40	24	13	3

BGVS



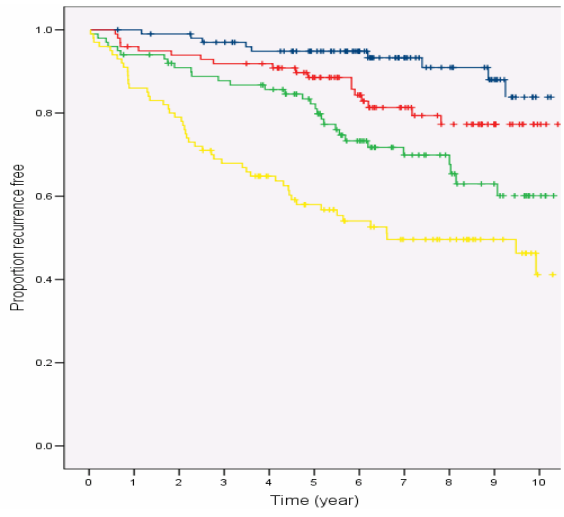
101	97	85	50	30	18
102	92	76	43	24	14
98	84	64	37	18	7
100	61	44	31	17	6

TSM



85	82	68	38	23	11
82	74	59	31	16	7
119	101	82	53	26	14
115	77	60	39	24	13

NPI^{q4}



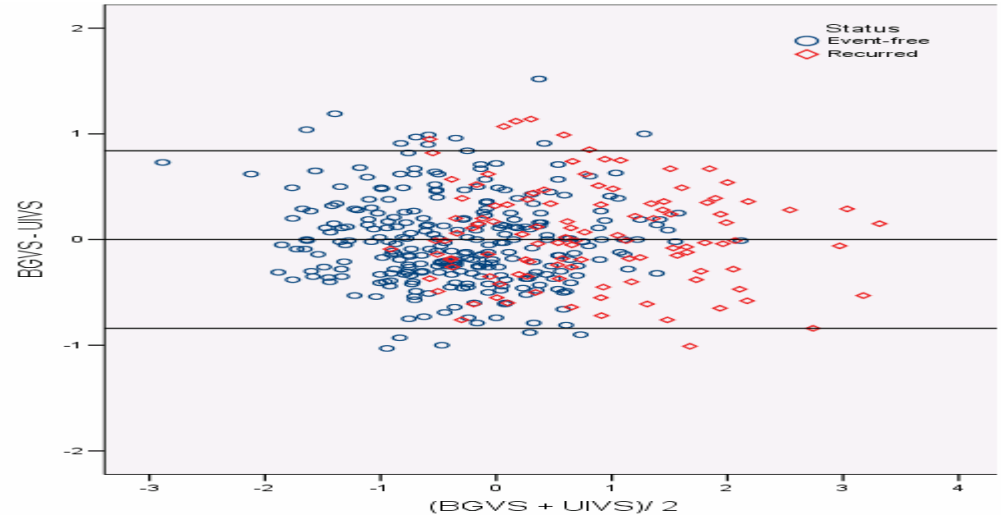
102	94	80	47	27	13
99	90	73	45	24	13
100	84	69	38	22	12
100	66	47	31	16	7

8.4.5 Comparison of the BGVS and UIVS approaches

i) Concordance between the BGVS and UIVS risk scores

Figure 8.6 plotted the mean of the BGVS and UIVS risk scores for each patient versus their difference with the 95% limits of agreement. There are no extreme outliers and no obvious patterns.

Figure 8.6: Assessing the agreement between the BGVS and UIVS risk scores



ii) Detection of low risk patients and estimated PSEP

As summarised in Table 8.4, there was a slight improvement in estimated RFS rates in the lowest risk group, this seen by integrating biological knowledge in the process of the model development. At 7-years the PSEPs were comparable (58% for the BGVS versus 55% for the UIVS) (Table 8.4), though at 10 years the BGVS risk groups gave better diverged risk groups (PSEP 71% versus 64%).

iii) Ability of the models to predict RFS in subset of low-risk patients

When excluding patients with NPI > 5.4, the sample size was reduced to 332. New quartiles for the BGVS and UIVS risk scores were calculated and applied to categorise patients into 4 risk groups. Actuarial 7-year RFS rate in the new BGVS and UIVS lowest risk groups was 99% and 95% respectively (Table 8.6). Since high risk patients were excluded, the ability to detect low risk group might not be interesting. However, new high risk groups with 7-year RFS around 50% were detected.

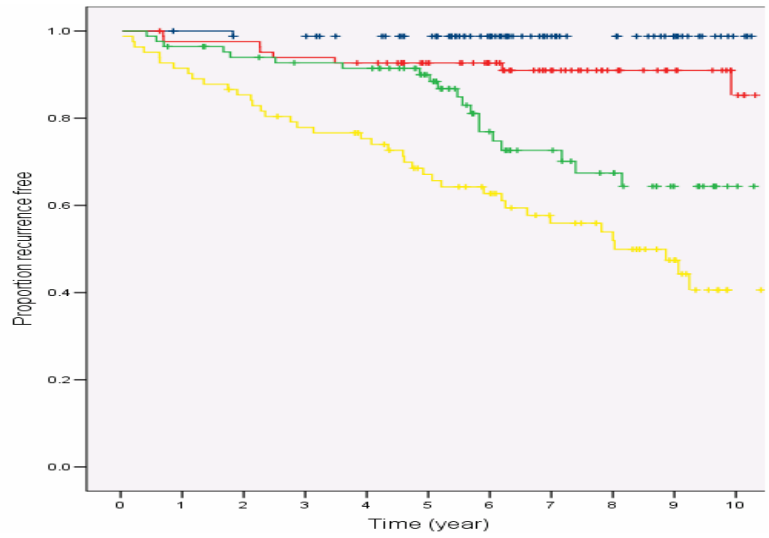
Both the new BGVS and UIVS risk groups divided the patients into four separate risk groups. PSEPs were 45% and 46%, respectively (Table 8.6 and Figure 8.7). When cut offs were applied at quartiles of NPI to categorise patients equally into 4 risk groups, estimated PSEP was only 26%.

Table 8.6: Ability of the BGVS and UIVS RFS risk groups to predict RFS in patients with NPI≤5.4

Risk group	Index	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	BGVS	99% (97%, 100%)	**99% (97%, 100%)**	99% (97%, 100%)
	UIVS	97% (93%, 100%)	**95% (89%, 100%)**	95% (89%, 100%)
Highest	BGVS	63% (51%, 75%)	**54% (42%, 66%)**	40% (26%, 54%)
	UIVS	60% (48%, 72%)	**49% (37%, 61%)**	40% (24%, 56%)
PSEP for BGVS risk groups		36%	**45%**	59%
PSEP for UIVS risk groups		37%	**46%**	55%

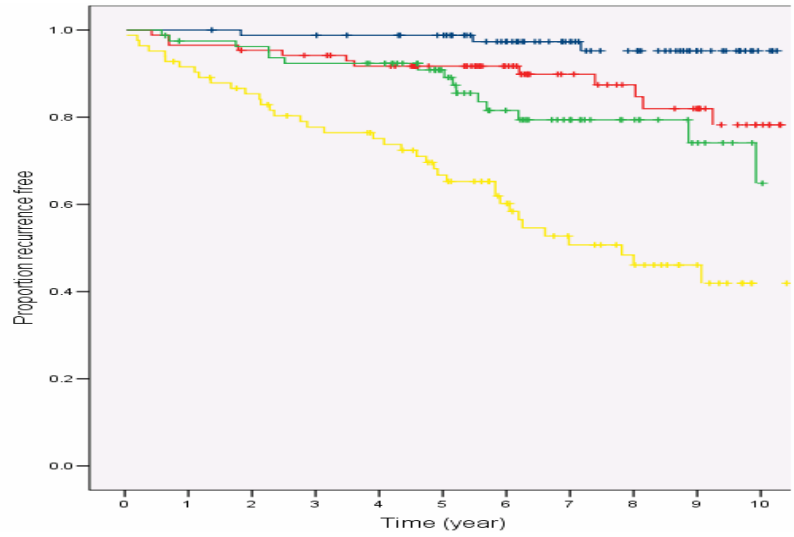
Figure 8.7: K-M curves applying the BGVS (top panel) and UIVS RFS risk groups (bottom panel) to predict RFS in subset of patients with NPI<= 5.4

BGVS



83	80	71	38	24	14
83	77	65	43	23	15
84	73	58	30	15	8
82	62	47	31	18	5

UIVS



83	81	76	50	28	16
86	79	66	38	27	16
80	72	53	29	13	7
83	60	46	25	12	3

8.5 Discussion

8.5.1 Form of association

In development of UIVS and BGVS Model, I retained maximum information by allowing for non-linear effects. In situations when external knowledge cannot guide the process of model development, as was the case here, application of data-driven model building strategies are required [Knorr KL et al., 1992; Harrell FE, 2001]. This is emphasized even when sample size is low [Marshall G et al., 1995].

In terms of the selection of an appropriate form and the predictive ability of the biomarkers, the current biological knowledge did not guide the process of model building. That is why a multi-faceted screening procedure was applied. There was therefore a risk of overfitting and chance influencing the models, these were then tackled via bootstrapping.

8.5.2 Imputation of missing data

Analysis of multiple biomarkers can be hampered by missing data, where samples are un-interpretable due to assay failure. Even a random loss of 1% of samples per assay could result in >30% of samples having missing data when 40 markers are evaluated, such as in the data set analysed for this research.

Even a low rate of missing data on each variable might cause serious problems in multivariate modelling when patients with missing data on different variables are not the same. This might substantially reduce the number of complete cases available for

analysis, and increase the chance of bias due to excluded cases. In order to protect against chance effects due to imputation I imputed 10 data sets, so that reliability across data sets (imputations) could be checked. This protection was felt to be worth the inconvenience of having to average risk scores across 10 final models.

As specified in Chapter 4 (section 4.4.4), in order to impute missing data for continuous biomarkers a Predictive Mean Matching Method was applied. Linear regression was not used to impute missing data for continuous variables as this method might produce out of range values. Although it would be possible to round the out of range values to the nearest boundary, this might result in a large proportion of imputed values being replaced by a single value (minimum or maximum) [Heitjan DF and Little RJA, 1991; Zhou XH et al., 2001].

8.5.3 The UIVS Model

A key step in the development of models when a large number of variables are available is to select a reduced set of variables prior to modelling. In development of the UIVS Model, in order to select informative biomarkers I allowed for non-linear effects, but then challenged the model obtained via bootstrapping to check both the stability of form, and reliability of inclusion across the variables. However, the UIVS Model required multiple comparisons, due to the application of statistical procedures to 72 TMA predictors.

The developed UIVS model, in comparison with NPI, found a noticeable improvement in model C-index and R-square. The actuarial 7-year RFS rate for the lowest risk group of NPI with 4 equal risk groups, and NPI with 3 risk groups, were

91% and 89%, respectively, while the corresponding figure for the UIVS was 95%. To detect a subset of low risk patients that do not require further treatments, such as adjuvant therapy, there is a need to shift a small number of recurred patients who are classified as low risk based on NPI, into intermediate or high risk groups. Therefore, these marginal improvements that were obtained might be very important in clinical practice. Furthermore, the UIVS risk groups were better diverged, yielding a 13 percentage point improvement in PSEP.

In a similar study, the ability of 126 antibodies to predict recurrence of breast cancer were assessed to develop a biomarker prognostic tool [Ring BZ et al., 2006]. In this the scoring of tissues was ordinal, and associations between each antibody and recurrence were assessed by applying a univariate Log-Rank test. In total, 20 biomarkers associated with outcome at a 0.10 significance level were used in a multifactorial Cox regression model. The final model comprised 5 biomarkers. Estimated 5-year non-recurrence rate in the lowest risk groups corresponding to NPI and biomarker models were 90% and 95%, respectively. The corresponding figures for the intermediate risk group were around 90% and 75%, respectively. This indicated that the biomarker model had better ability to select low risk patients, and also to stratify patients into better diverged risk groups. However, I feel that antibodies have a continuous nature, the authors applied an ordinal scaling but have not described the nature of initial distribution and rationale of categorisation. No attempt was made to optimise the form of association for each biomarker.

8.5.4 Check of stability of transformations

In development of the UIVS Model, 5 biomarkers with univariate non-linear risk functions (3 polynomials, 1 threshold, and 1 non-ordinal) were offered to the multifactorial model. The stability of transformations found in screening phase was checked across 1000 bootstrap samples. There was no contribution of polynomial effects of Rkipnu and Ptennu in the UIVS Model, this is because the form of association was not consistent across 1000 bootstrap samples (see Chapter 8 for details).

Although the stability was checked across 1000 samples, to reduce the burden of model building the results were aggregated across 10 imputed data sets. Therefore, I also investigated whether it was enough to check the stability of transformations across only 10 imputed data sets. By chance results were exactly the same as before.

As a third and much easier option, I simply applied the transformations to the biomarkers and applied the transformed versions to the multifactorial model. In this case, both of Rkipnu and Ptennu were forced into the multifactorial model (data not shown). This highlighted the importance of investigating the stability of forms, and refining the models by exclusion of unreliable predictors, in order to avoid unstable prognostic models.

8.5.5 The BGVS Model

When many variables are available, incorporation of biological knowledge in the process of model development is very important, and it plays an important role in reduction of variables. It has been noted that predictors already reported as prognostic factors should normally be candidates for a multifactorial model. Furthermore, predictors which are highly correlated with other variables should be omitted as they likely contain little extra prognostic information [Harrell FE, 2001].

Biomarkers were evaluated by analysis of the cell cytoplasm, nuclei, and membranes from each sample, there were therefore correlations between expression of biomarkers from each these three sources. Reported correlations were higher than 50%, so I explained this issue to the biologists and asked them to exclude one of biomarkers which they think might be more difficult to measure or does not have prognostic value. However, they explained that 50% correlation is not high from a biological perspective. In addition, they collected data from all 3 cell components to understand better the biological process of cancer recurrence, they therefore did not exclude any biomarkers.

In development of the BGVS Model Bayesian methods were used implicitly, as the biological knowledge available about tumour progression pathways was used to define substantive family sets. Using biological knowledge, family risk scores were used in the development of the prognostic model. I decided to calculate an index that was representative of that set in the multifactorial model (risk score). One approach was to give weight to variables reflecting their biological importance, however

information on appropriate weighting was not available. There would be scope for further enhancement of the BGVS Model as knowledge of the role of individual proteins within families develops.

For the BGVS model, in comparison with NPI, a noticeable improvement in C-index (79% versus 72%, respectively) was seen, and R-square was doubled. For the 25% of patients with the lowest prognostic risk scores the estimated 7-year RFS rate was 98% with BGVS, this is much higher than the 79% with the conventional NPI^{std3} classification. The BGVS risk grouping significantly reclassified nearly 21% of recurred and 11% of non-recurred subjects into a more appropriate risk group.

The BGVS Model might suffer overfitting since the model was developed by applying the MFP on composite scores preceding the multifactorial Cox or MFP models.

8.5.6 Ability of the UIVS and BGVS Models to predict other end points

One of the clinical questions posed by Professor John Bartlett was whether the UIVS and BGVS Models developed to predict RFS can predict other end points. From a statistical point of view it might be strange to develop a model for RFS, and to then investigate whether it can stratify patients into risk groups with respect to another outcome. However, Professor John Bartlett explained that whilst the primary end

point is RFS, the clinical community will want to see whether it can be used to predict OS, as changes in RFS do not always translate into changes in OS.
Results presented in Appendix 1 indicate the ability of both the UIVS and BGVS Models to stratify patients into risk groups with respect to Recurrence Free on Tamoxifen (RFoT) and Overall Survival (OS).

A separate model was not developed for OS as there were only 74 deaths and therefore the results would not be robust. However, applying the methods explained in this section, prognostic models were developed using RFoT as the outcome. Those models are not reported in this thesis but are given in a manuscript for publication.

One of the interesting differences between the RFS and RFoT models, in terms of biomarkers retained in the model, was that HER2 contributed only to the RFoT model. On the other hand, the biomarkers Akt1nu and Tunel were predictors for RFS but not RFoT. Neither of these variables (Akt1nu and Tunel) was selected in the screening phase to be applied to RFoT model. For the RFoT model a smaller number of events were analysed (n=84) so the power was lower than in the RFS model. Despite the model refinement techniques employed, this lower power might explain some of difference in terms of biomarkers included in the final model.

8.5.7 Tree-based Survival Model (TSM)

Although there is no evidence that TSM produce better models than standard regression methods, this model is frequently used due to its novelty and simplicity. TSM, in contrast with the UIVS and BGVS approaches, was very simple as neither

data-reduction nor missing value imputation was required. Assessment of form was also irrelevant since this method dichotomises the biomarkers into low versus high risk groups. TSM found a low risk group of 84 patients with 3 recurrences. Since this method created risk groups with different sizes, in comparison with other methods, no further exploration was done.

The main weaknesses of the techniques used are multiple testing and overfitting [Clark TG et al., 2003]. It has been discussed that trees are sensitive to even small changes in the sample [Lausen B et al., 2004]. Tree-based methods usually find prediction rules that validate poorly, due to exhaustive searching to select covariates and their cut points [Harrell FE et al., 1998; Clark TG et al., 2003]. However, TSM analysis involves no preliminary steps and can be used as a good approximation for a complex model.

To improve the prediction of probability of survival, bagging of survival trees is proposed. In this approach, by re-sampling from the original data a large number of trees is constructed. The aggregated Kaplan-Meier curve for a new patient is defined as the Kaplan-Meier curve of all observations identified by the M leaves containing the new patient [Hothorn T et al., 2004]. Therefore, no single tree can be reported and communication of results is not simple.

8.6 Overview

The performance of biomarker models was found to be superior to NPI in terms of risk classification, and selection of low and high risk patients. This was determined by performing 3 modelling procedures to detect the form of association, followed by imputation of missing data, and checking the stability of the models. The number of biomarkers that contributed significantly in the UIVS and BGVS Models were 6 and 8, respectively. On the other hand, only 3 biomarkers contributed to the decision tree. Two clinical variables (tumour size and nodal stage) were retained in all 3 models. Prhisto was in the TSM but was not retain in regression models. In the coming chapters I will enhance the understanding of these processes with investigation of elements of the procedures.

8.7 Chapter summary

- When I allowed for non-linear effects (UIVS Model), significant improvement over NPI was seen. Risk groups were better diverged, a significant proportion of patients were allocated into a more appropriate risk group, and a considerable reduction in number of events observed in the lowest risk group was seen.

- Incorporation of biological knowledge slightly enhanced model performance in terms of selection of low-risk patients, classification of patients into more appropriate risk groups, and ability to predict other end points.

- TSM was a good approximation for complex regression models

Chapter 9 EXAMINATION OF METHODS APPLIED: RELAXING THRESHOLD FOR INCLUSION OF VARIABLES IN THE MULTIFACTORIAL MODEL AND COMPARING IMPUTATION METHODS

9.1 Introduction

In the screening for the univariately informative biomarkers (Chapter 7), the P-value threshold to declare a biomarker as being informative was set at 0.10. This was selected to reduce the number of variables to be applied to the model, in order to avoid convergence problems, but at the same time avoid missing any biomarkers that might be important in predicting the outcome.

In this chapter, I explored the effect of relaxing the 0.10 P-value threshold used to select informative variables, both in terms of whether it might cause convergence problems for the MICE method (to impute missing values) [Heymans MW et al., 2007], and also to explore the composition of the multifactorial model when univariately insignificant biomarkers are submitted to the model. One solution to such a situation would be to substitute the MICE imputation with replacing missing data with median values, since this runs no risk of lack of convergence. Furthermore, it has been reported that, this method is a good approximation for the MICE when there are few missing values [Shrive FM et al., 2006; Musil CM et al., 2002; Barzi F and Woodward M, 2004].

9.2 Aim

The main aims of this part of the research are to:

1. Assess the impact of replacing missing data with median values on performance of the UIVS Model

2. Compare the ability of the MICE and 'Median Substitution' imputation techniques to impute the missing data when many potential variables are available

9.3 Methods

All models presented in this chapter were developed in conjunction with Backward Elimination (B.E.) variable selection method, and no bootstrap resampling procedure was applied.

9.3.1 Replacement of missing data by median in the UIVS Model

The UIVS Model presented in Chapter 8 was redeveloped, in this case instead of using the MICE method [Van Buuren S et al., 1999], missing data were replaced by the median of observed values. The UIVS Model (presented in Chapter 8) and a new model, which is called the UIVS* Model, were compared in terms of selection of variables and estimated S.E.'s, and statistics explained in section 4.4.7.

9.3.2 Process of development of the multifactorial models

In total 6 models were developed, this was done by using different P-values to declare biomarkers as being informative in the screening phase (3 different P-values) and by applying alternative imputation methods (2 methods). Standard Errors (S.E.) derived from the UIVS Model (see Chapter 8) were taken as gold standard. I recorded S.E. of variables retained in the UIVS Model under 6 different scenarios. Estimates were presented graphically.

i) Relaxation of the 0.10 P-value threshold to declare a biomarker as being informative

First biomarkers with a univariate P-value of less than 30% were offered to the multifactorial model. Then, this threshold was further relaxed to 50%. Finally, all biomarkers, regardless of their univariate P-value, were used. Clinical variables were used in all scenarios.

ii) Imputation of missing data

Missing data were replaced by 10 imputed values applying the MICE method. Alternatively missing values were substituted by the median of observed values.

iii) Development of multifactorial models

When the MICE method was used to impute missing data (section 4.4.4), the stability of the threshold effect of Pmapknu, and the non-ordinal effect of Akt1nu, was checked across 10 imputed data sets. Transformation was applied if replicated in more than 5 samples. Then, Multivariate Fractional Polynomial (MFP) was applied (section 4.4.2) to each of the 10 data sets to check whether power transformation of the biomarkers showing univariate polynomial association (Krascy, Rkipnu, and Ptennu) with outcome (Recurrence Free Survival (RFS)) improves the fit. Transformation was applied if optimum form (FP2, FP1, linear) was replicated in more than 5 samples. Estimates were aggregated as explained in section 4.4.4 parts i and ii. When missing data were replaced by the median of observed values, the threshold effect of Pmapknu, and non-ordinal effect of Akt1nu was applied. This is

because only 1 data set was imputed. To develop the multifactorial model, MFP was applied.

9.4 Results

9.4.1 Impact of the imputation method on performance of the UIVS Model

The estimated 95% confidence intervals of the Hazard Ratios (HR) are summarised in Table 9.1. Variables retained in the multifactorial models were the same. Omitting bootstrap procedure, by chance, and none of the biomarkers with unstable form were retained in the multifactorial model. The resulting estimated HR's and S.E.'s were comparable (Table 9.1). In terms of performance, replacement of missing data by median resulted in a 2 percentage point reduction in C-index (Table 9.1), as well as a slight reduction in predictive ability and goodness of fit.

Risk groups derived from two imputation methods applied were compared with NPI with 4 equal risk groups (NPI^{q4}). A slight reduction in estimated Net Reclassification Index (NR Index) was seen (18% for the MICE versus 16% for the median substitution). On the other hand, estimated RFS rates in the lowest and highest quartiles of the MICE and median substitution methods and PSEPs were similar (55% versus 53%) (Table 9.2). K-M curves corresponding to risk groups derived are given in Figure 9.1. It can be seen that while the lowest and highest risk groups remained fairly similar, middle risk groups corresponding to the MICE method were better diverged.

I also checked the distribution of patients into risk groups based on the risk groups that were derived (see Appendix 2). The percentage of recurred and non-recurred patients that were located in the same risk group was 72% and 69%, respectively. Out of 112 recurred patients, risk groups derived applying the MICE method, relative to the median substitution method, shifted 16 into a more appropriate, and 13 patients into a less appropriate risk group. The corresponding figures for 289 non-recurred patients were 48 and 42, respectively. This indicated that there were no significant differences between models in terms of assignment of patients into risk groups.

Table 9.1: Comparison between the MICE and median substitution methods on estimated HR's and S.E.'s in the UIVS Model

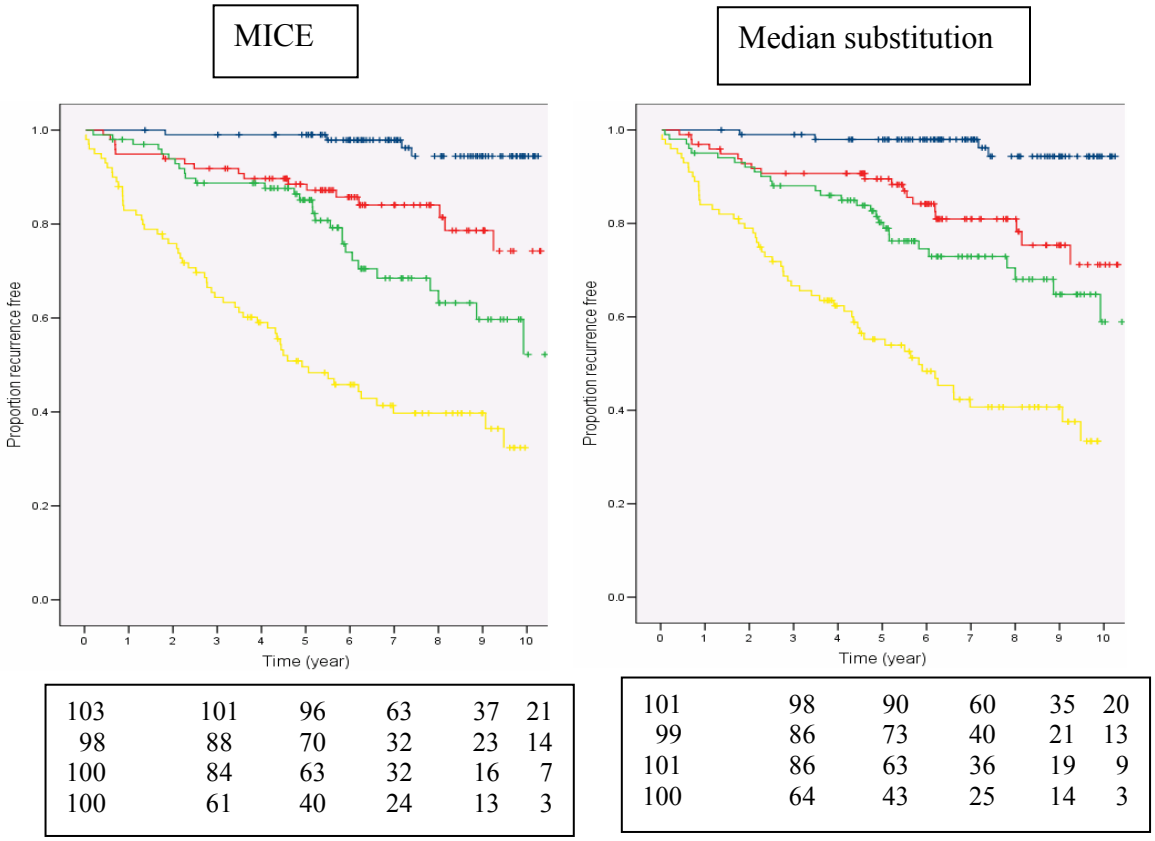
Variable	MICE method (Chapter 8)			Median substitution method		
	HR (95% C.I.)	P-value	S.E.	HR (95% C.I.)	P-value	S.E.
Nodal	1.82 (1.38, 2.40)	<0.001	0.14	1.61 (1.25, 2.06)	<0.001	0.13
Size (cm)	1.21 (1.10, 1.31)	0.001	0.05	1.22 (1.11, 1.32)	<0.001	0.05
Pmtor	0.33 (0.19, 0.59)	<0.001	0.28	0.33 (0.20, 0.61)	<0.001	0.28
Tunel	1.49 (1.23, 1.81)	<0.001	0.11	1.49 (1.22, 1.82)	<0.001	0.11
Praf338cy	2.12 (1.07, 4.02)	0.03	0.34	2.45 (1.22, 4.89)	0.01	0.33
Pmapknu	2.80 (1.72, 4.57)	<0.001	0.25	2.77 (1.72, 4.47)	<0.001	0.24
Krascy	6.05 (2.23, 16.44)	<0.001	0.51	7.84 (2.78, 22.17)	<0.001	0.53
Akt1nu	0.54 (0.36, 0.82)	<0.001	0.21	0.53 (0.35, 0.79)	0.002	0.21
Performance						
C-index	79%			77%		
R-square	28%			25%		
Chi-square	123.6			108.8		
NR Index* (P-value)	18% (0.02)			16% (0.04)		

*Relative to NPI with 4 equal risk groups

Table 9.2: RFS rates in the lowest and highest quartiles of the UIVS risk scores applying alternative imputation methods

Risk group	Imputation method	5-year event free (95% C.I.)	7-year event free (95% C.I.)
Lowest	MICE	98% (97%, 100%)	**95% (89%, 100%)**
	Median substitution	98% (96%, 100%)	**94% (88%, 100%)**
Highest	MICE	46% (36%, 56%)	**40% (30%, 50%)**
	Median substitution	48% (38%, 58%)	**41% (31%, 51%)**
PSEP for MICE method		52%	**55%**
PSEP for Median substitution method		50%	**53%**

Figure 9.1: K-M curves for the UIVS risk groups applying MICE (left panel) and Median substitution imputation methods (right panel)



9.4.2 Impact of relaxation of the 10% P-value threshold and imputation method on the composition of the UIVS Model

i) Modelling variables with univariate P-value less than 30%

A list of all 72 biomarkers with univariate P-values in univariate Fractional Polynomial (FP) analysis is given in Appendix 3. Thirty two variables (about 45% of variables) were candidates for this phase of analysis. This means the EPV is 3.5.

While overall only 6.8% of values of all variables were missing, only 168 patients had complete data on all 32 variables. In the multifactorial analysis, applying the MICE method, the only difference with the UIVS Model (presented in Chapter 8), was the inclusion of Praf259cy instead of Praf338cy (Table 9.3). Replacement of missing data by median values gave similar results, except that Jrh3me (with 40% missing values) was retained in the model with a P-value of 0.02.

ii) Modelling variables with univariate P-value less than 50%

A total of 47 variables (65% of available covariates) were candidates to be applied to the multifactorial models, this gave an EPV of 2.4. In total, 238 patients had at least one missing datum on any of 47 variables. Overall only 6.6% of values of all variables were missing.

In the multifactorial analysis, regardless of the imputation method used, the results were either the same and similar to that of 30% threshold value. None of the biomarkers with a univariate P-value between 30% and 50% were retained in the multifactorial model.

iii) Modelling all variables

Applying all 72 variables to the model gave an EPV of 1.6. Only 126 patients had complete data on all variables, and the proportion of missing value for the all values of all variables was as low as 7.4%.

When the MICE method was used, results were either the same and similar to that of 30% threshold value. However, when missing data were replaced with median values, Akt1nu was removed, while Jrh3me was retained in the model (with P-value of 0.02) (Table 9.3).

Table 9.3: Modelling variables with univariate P-value < 30%: inclusion of variables in the multifactorial models applying different imputation methods

Variable	Number of patients with missing data	Univariate P-value	MICE	Median replacement
Nodal	33	<0.001	Yes	Yes
Size	22	<0.001	Yes	Yes
Krascy	14	<0.001	Yes	Yes
Akt1nu	5	0.003	Yes	Yes
Pmapknu	33	0.003	Yes	Yes
Pmtor	11	0.02	Yes	Yes
Tunel	39	0.07	Yes	Yes
Praf259cy	35	0.20	Yes	Yes
Jrh3me	166	0.27		Yes

When P-value relaxed to 50% results were exactly the same
When all variables were submitted to the model, in median substitution method Akt1nu lost its effect

9.4.3 Investigation of inflation of S.E.s

The S.E.'s for variables contributing to the UIVS Model when different variables were applied to the multifactorial models, and alternative imputation methods were used, are plotted in Figure 9.2. S.E.'s reported are those from the first step of modelling because with application of the B.E. variable selection method Praf338cy was not retained in the models.

For continuous biomarkers, the S.E. of regression coefficients corresponding to a 100 unit change in the biomarker is reported. As the P-value threshold increased (as EPV therefore decreased) the S.E.s increased. Estimated S.E.'s in the MICE model was slightly larger than in the 'Median Substitution' model. The differences between estimated S.E.'s when alternative imputation methods were used are plotted in Figure 9.3. It can be seen that as EPV decreased, the difference between S.E.'s became larger.

Figure 9.2: Estimated S.E.'s for variables retained in the UIVS Model at different threshold P-values by applying MICE method

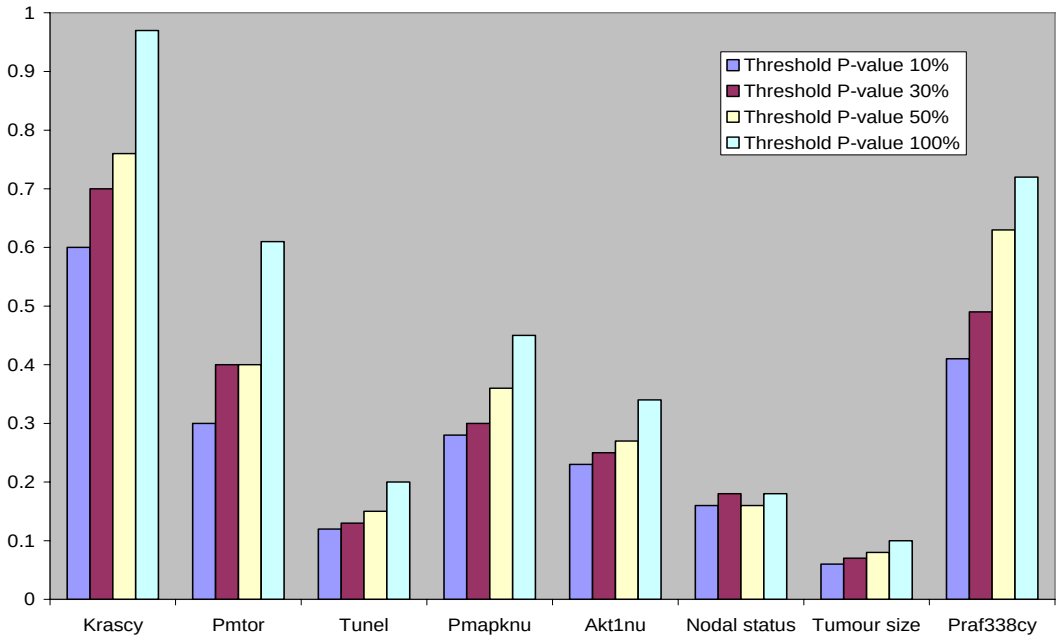
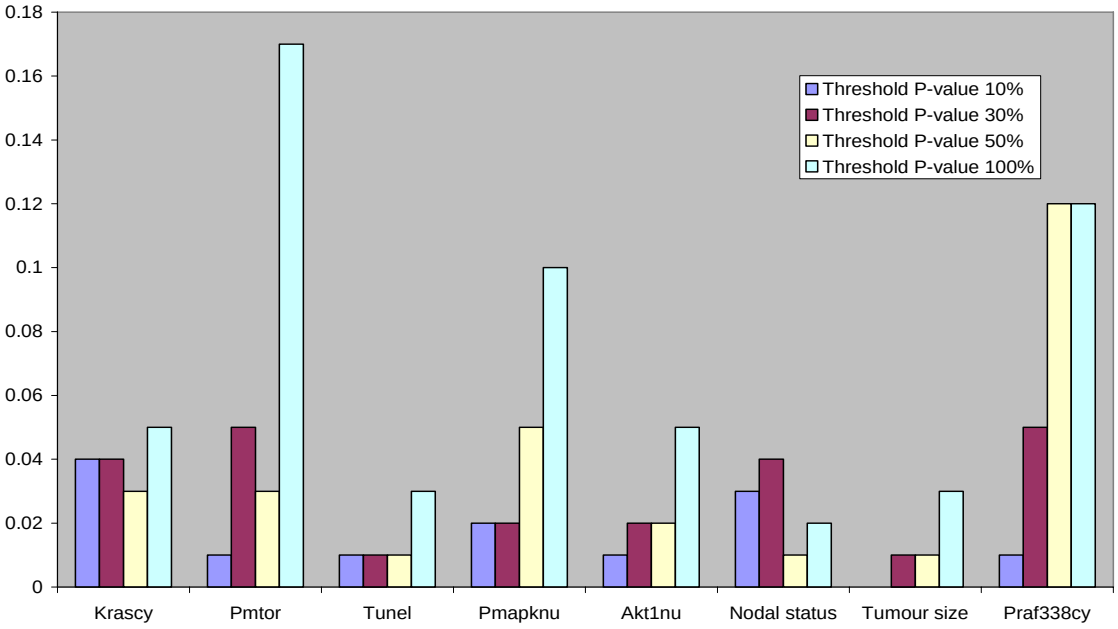


Figure 9.3: Difference of estimated S.E.'s for variables retained in the UIVS Model, at different threshold P-values, applying MICE and Median substitution imputation methods



9.5 Discussion

The results presented show that when applying the MICE imputation method, inclusion of univariately non-significant variables made a slight change in the composition of the UIVS Model. Only praf259cy was retained instead of Praf338cy.

An important issue is that applying many variables to the MICE imputation methods might cause convergence problems. This happened in Heymans study, in that the imputation model did not converge when 31 variables were applied to MICE [Heymans MW et al., 2007]. I did not have a convergence problem with this dataset but it might happen in other data sets.

Replacement of missing data by the median of observed values artificially reduces the variance of the variables containing missing data [Croy CD and Novins DK, 2005]. However, it is a good approximation for sophisticated methods such as the MICE [Van Der Heijden GJ et al., 2006; Barzi F and Woodward M, 2004; Kristman VL et al., 2005]. When I used this approach and compared UIVS and UIVS* Models, I saw that the variables that were retained in the models were the same, but performance of the UIVS Model was slightly better. Furthermore, when I submitted univariately non-significant variables to the model, slight differences in composition of the UIVS Model was seen, in that Praf259cy and Jrh3me were then retained in the models (see Table 9.3 for details).

'The optimal method should balance validity, ease of interpretability for readers, and analysis expertise of the research team' [Shrive FM et al., 2006]. The MICE method

does not provide unique estimates [Kneipp SM and McIntosh M, 2001]. Furthermore, to apply the technique, special software is required and communication of results with clinical audiences is not simple. 'Median Substitution', on the other hand, is an easy and fast method which can be applied simply and communicated with clinicians. Furthermore, this method is deterministic and gives unique results but artificially reduces the variance.

In my opinion, median substitution is a reasonable approach when both the missing rate for an individual variable, and number of cells with missing values in the whole data set is low. As an example, consider a data set containing 8 predictors, survival time, and outcome, with sample size 400 (4000 cells in total). If the value of one of the variables is missing for 200 patients, then the missing rate for that specific variable is 50% while the proportion of cells with missing values is low at 0.05% (200 / (8+2) x 400). Having a low proportion of cells with missing data in the whole data set might not guarantee suitability of replacement by the median method. Although approaches such as mean or median imputation might give results comparable to the MICE, in terms of variables contribute to the multifactorial model, a gold standard (MICE) is required to compare results from other simpler methods [Greenland and Finkle, 1995].

9.6 Chapter summary

- Modelling informative variables, the 'Median Substitution' imputation method gave results comparable to the MICE. In our data set this might be due to a low rate of missing data. However, substitution of missing data by a single value artificially reduces the variance of the variable.

- When I relaxed the 10% P-value threshold and included univariately non-significant biomarkers into the model, the structure of the UIVS Models was very slightly changed in particular when missing data were replaced by median.

Chapter 10 EXAMINATION OF METHODS APPLIED: EXPLORING FORM SCREENING METHODS

10.1 Introduction and the background

The methodology applied in Chapter 8 (to develop the UIVS Model) was not easy to apply nor is it straightforward to communicate the results with clinical audiences. In Chapter 9, to reduce the complexity the bootstrap step was omitted and missing data were replace by the median of observed values (the UIVS* Model). Omission of the bootstrap step and substitution of missing data by median of observed values hugely simplified the process of model building, and none of the biomarkers with unstable non-linear effect (Rkipnu and Ptennu, see Chapter 8 for more details) were retained in the model by chance.

However, the approach utilised still has a complicated screening stage. This involved the application of many univariate tests which might increase the possibility of false positives due to the multiple comparisons undertaken. My approach also differed from the usual model building practice where the researcher generally applies only a single method, usually linear Cox model, due to the availability of user friendly software such as SPSS.

I demonstrated that the UIVS Model had considerably better performance than Nottingham Prognostic Index (NPI). I wished to investigate whether the better performance seen was attributed to the exhaustive screening procedure applied.

Another issue is that, in development of the UIVS Model, my priority was to keep the biomarkers in continuous form. That was because Royston *et al.* showed that dichotomisation of continuous data, and the inevitable loss of information ensuing, might affect the performance by diminishing the goodness of fit, discrimination, and predictive ability [Royston P et al., 2006]. I therefore, by applying both data-driven and pre-specified methods of variable selection, wished to investigate whether the same conclusion was true in highly skewed data, such as the current breast cancer data set I analysed.

10.2 Aim

The main aims of this part of the research are to:

1. Compare the performance of multifactorial models in which biomarkers were kept in continuous form or transformed into dichotomised form

2. Compare the performance of data-driven and pre-specified screening methods (to select informative biomarkers and their form) on performance of the models

3. Compare the performance of all models with UIVS, in order to address whether the better performance seen over NPI was due to complex screening procedure applied

10.3 Methods

The results of screening methods to select the biomarkers and their form were used, these are presented in Chapter 7 (see section 7.3 for details of methods, and Tables 7.2 and 7.3 for results). Clinical variables (nodal status, grade, tumour size) were used in development of all models, and the MICE method was applied to impute missing data (section 4.4.4). Multifactorial models were developed in conjunction with B.E., aggregation of the results and a comparison of models was performed as explained in section 4.4.4 part ii and section 4.4.7.

10.3.1 Keeping the biomarkers in continuous form

i) Data-driven selection by Fractional Polynomial (FP) model followed by multifactorial modelling using MFP model

FP was applied to select biomarkers with univariate P-value < 0.10, followed by application of Multivariate Fractional Polynomial (MFP) (section 4.4.2). FP explores a range of power transformation to estimate optimised power(s). Through this chapter, this model was named MFP model (Table 10.1).

ii) Pre-specified selection by linear Cox model followed by multifactorial modelling using linear Cox model

To develop this model, I assumed that a prior linear form was adequate to represent the effect of skewed biomarkers. Univariate Cox was applied to select biomarkers which were significant at a 0.10 levels. Multifactorial linear Cox model was then fitted. This model was named 'linear Cox model'.

10.3.2 Dichotomisation of biomarkers

Three methods were applied to dichotomise the biomarkers, and the resulting binary version of biomarkers selected was used in the multifactorial binary Cox model (Table 10.1).

i) Data-driven split selection by Minimum P-value method

Minimum P-value method was applied to find the split optimised for data (section 4.4.3). To avoid groups with small number of patients, I did not apply split at the outer 20% of distribution of biomarkers. Biomarkers with P-value < 0.005 were candidate for this model (equivalent to a 0.10 in a linear Cox model) [Altman DG et al., 1994]. This model was named 'Optimal split model'.

ii) Data-driven split selection by dichotomisation at one of the quartiles

A second less extreme data-driven model was developed. To choose the appropriate split, I decided to apply the split in turn at first, second, and third quartiles and then selected the split to be used at which the highest number of biomarkers were significant at a 0.033 level (equivalent to a 0.10 in a linear Cox model). This model was named 'Quartile model'.

iii) Pre-specified split selection (median)

By pre-specifying the split at median, I made the place of split blind to the data. Biomarker variables with univariate P-value < 0.10 were selected. This model was named 'Median model'.

Table 10.1: Methods used to screen biomarkers and to detect the appropriate form of risk function

Type of biomarker	How form/ place of split was selected	Nature of univariate screening method applied	Multifactorial analysis	Model name through this Chapter
Continuous	Optimised power transformation	Data-dependent	MFP	MFP model
	Linear form	Pre-specified	Linear Cox model	Linear Cox model
Dichotomised	Optimal split	Data-dependent	Binary Cox model	Optimal split model
	Quartile at which higher number of biomarkers are significant	Data-dependent	Binary Cox model	Quartile model
	Split at median	Pre-specified	Binary Cox model	Median model

10.4 Results

10.4.1 Biomarkers candidate for multifactorial models

The biomarkers and forms selected to be used in the multifactorial models are listed in Table 10.2. To develop each of the models described, the biomarkers with bold P-values plus clinical variables were used.

By applying the optimal split method the cut point selected for Rkipnu was 8, with P-value of 0.001. However, Professor John Bartlett explained that it would be very difficult in practice to distinguish patients by applying this low histoscore value. Therefore, this biomarker was not offered to the Optimal split model (Table 10.2).

Following results reported in Chapter 7, to develop the Quartile model the split was applied at the upper quartile, this was because 8 biomarkers were significant at this split. The number of significant biomarkers when applying the split at lower quartile and median was 2 and 4, respectively. These biomarkers were also selected by applying the split at top quartile (see details in chapter 7).

Akt1nu was used in development of UIVS Model but was not used in development of models presented here, this is because none of methods considered here selected this biomarker as informative.

Table 10.2: Biomarkers screened with different methods

Row	Screening method	Continuous form		Binary form by dichotomisation at		
		FP	Linear Cox	Optimal split	Upper quartile	Median
	Biomarkers	P-value	P-value	Optimal split (P-value)	Upper quartile (P-value)	Median (P-value)
1	Praf338cy	**0.01**	**0.01**	**192 (0.001)**	**190 (0.001)**	**167 (0.01)**
2	Praf338nu	**0.002**	**0.002**	**123 (0.001)**	**158 (0.01)**	**135 (0.01)**
3	Prhisto	**0.007**	**0.007**	**20 (0.001)**	**140 (0.02)**	**35 (0.001)**
4	Akt2cy	**0.06**	**0.06**	**190 (0.005)**	**188 (0.01)**	**158 (0.03)**
5	Mapkcy	**0.01**	**0.01**	**128 (0.003)**	**147 (0.03)**	110 (0.11)
6	Pmtor	**0.02**	**0.02**	**100 (0.001)**	**90 (0.02)**	50 (0.40)
7	Tunel	**0.07**	**0.07**	**105 (0.003)**	**72 (0.02)**	0 (0.54)
8	Pher2nu	**0.07**	**0.07**	**80 (0.005)**	65 (0.33)	**43 (0.07)**
9	Mtor	**0.06**	**0.06**	127 (0.01)	**105 (0.024)**	65 (0.17)
10	Krascy	**<0.001***	0.59	7 (0.01)	85 (0.53)	53 (0.72)
11	Rkipnu	**<0.001***	0.65	*8 (0.001)*	50 (0.59)	28 (0.65)
12	Ptennu	**0.02***	0.83	2.5 (0.01)	53 (0.66)	25 (0.80)
13	Pmapknu	0.92	0.92	**104 (0.003)**	95 (0.15)	72 (0.99)
14	Tescy	0.12	0.12	120 (0.03)	170 (0.10)	**112 (0.08)**

Threshold P-values for selection of informative biomarkers were: 0.10 in FP, linear Cox, and median models, 0.005 in minimum P-value to detect optimal split, and 0.033 in quartile dichotomisation methods
* Polynomial form

10.4.2 Continuous biomarker models

i) Comparison between data-driven MFP and pre-specified linear Cox models

The number of biomarkers used in development of MFP and linear Cox models was 12 and 9, respectively (Table 10.2), of which 6 and 5 variables, respectively, significantly contributed to the multifactorial models (Table 10.3).

The main difference between the MFP and linear Cox models was the contribution of Krascy, which had a polynomial effect in the MFP model. An FP2 transformation captured the effect of this biomarker, with the best powers across all 10 data sets being (3, 3). Inclusion of Krascy resulted in an increase in discrimination ability (C-index), from 73.5% to 75.5% (Table 10.3). In addition, an improvement was observed in estimated 7-year RFS in the lowest-risk group (94% versus 90%) and estimated PSEP (54% versus 45%), indicating greater ability of MFP index to distinguish low and high risk patients (Table 10.4).

It appears that performance of the MFP model was superior to linear Cox model. K-M survival curves corresponding to risk groups derived from MFP and linear Cox indices are given in Figure 10.1. High risk patients detected by MFP exhibited a worse recurrence free experience than high risk patients detected by linear Cox model.

ii) Comparison of continuous biomarker models (MFP and linear Cox) with NPI

Figure 10.2 (bottom panel) showed that the linear Cox risk groups and NPI^{q4} classified 53% and 44% of recurred and non-recurred patients into the same risk

groups. Corresponding rates for MFP risk groups were 54.4% and 45.3% respectively (top panel). Linear Cox and MFP risk groups tended to be less likely to classify recurrence patients into the lower risk group than NPI^{q4}. It also was more likely to classify non-recurred patients into lower risk groups than NPI^{q4}.

Estimated 7-year RFS of lowest risk groups corresponding to MFP and NPI risk models were 94% and 91%, respectively (Table 10.4). MFP risk groups, in comparison with NPI^{q4}, shifted 29% of recurred patients into a more appropriate and 17% into a less appropriate risk group (see Appendix 2). This gave a net gain of 12 percentage points. Corresponding figures for non-recurred patients were 30% and 25%, respectively, with net gain of 5 percentage points. Estimated Net Reclassification Index (NR Index) was 17% (P-value=0.02).

Comparing linear Cox risk groups with NPI^{q4}, the estimated 7-year RFS rates were similar (Table 10.4). In total, 26% of recurred patients were allocated into a higher risk group while 21% into a lower risk group giving net gain of 5 percentage points (see Appendix 2). Corresponding figures for non-recurred patients were 26% and 29%, respectively, giving improvement of 3 percentage points. NR Index was 8% which was not significant (P-value=0.34).

Therefore, it appears that integration of biomarkers with a linear risk function (linear Cox model) did not improve the ability of NPI in terms of detection of low risk patients, or risk group assignment. On the other hand, application of data-driven FP and optimisation of power transformation improve both features.

Table 10.3: MFP versus linear Cox model: comparison of hazard ratios, and performance of indices

Model	MFP model		Linear Cox model	
	HR (95% C.I.)	P-value	HR (95% C.I.)	P-value
Nodal	1.78 (1.35, 2.37)	<0.001	1.90 (1.45, 2.48)	<0.001
Size (cm)	1.20 (1.10, 1.30)	0.001	1.20 (1.10, 1.30)	<0.001
Pmtor	0.37 (0.20, 0.66)	<0.001	0.43 (0.26, 0.72)	<0.001
Tunel	1.49 (1.23, 1.81)	<0.001	1.36 (1.09, 1.68)	0.006
Praf338cy	2.22 (1.06, 4.69)	0.03	1.88 (1.08, 3.23)	0.04
Krascy	7.10 (2.37, 21.27)	<0.001	Not screened in	
Performance				
C-index	75.5%		73.5%	
R-square	21%		18%	
Chi-square	88.6		78	
NR Index (P-value)	17% (0.02)		8% (0.34)	

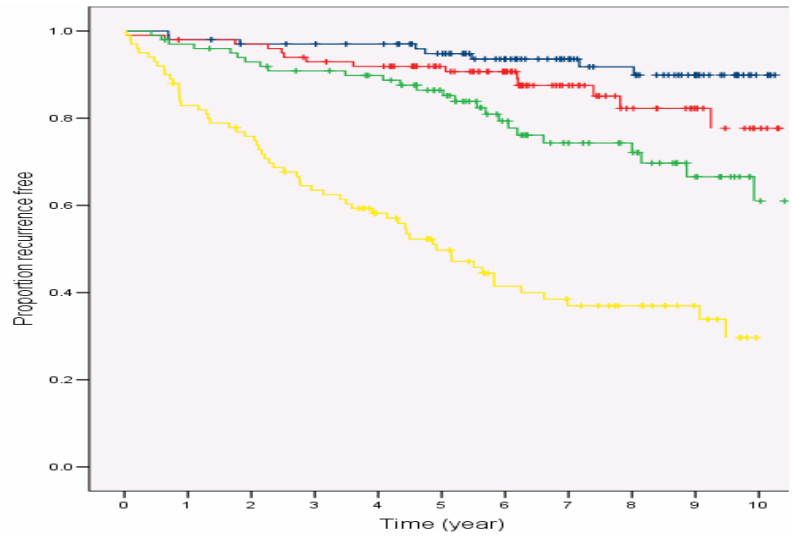
For biomarkers, HR shows amount of increase in risk of recurrence per 100 unit change in the value of the predictors

Table 10.4: Estimated RFS rates in the lowest and highest quartiles of MFP and linear Cox indices

Risk group	Index	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	MFP	97% (93%, 100%)	**94% (88%, 100%)**	80% (62%, 98%)
	Linear Cox	93% (87%, 99%)	**90% (84%, 96%)**	82% (70%, 94%)
	NPI	95% (91%, 99%)	**91% (85%, 97%)**	84% (72%, 96%)
Highest	MFP	44% (34%, 54%)	**40% (30%, 50%)**	30% (18%, 24%)
	Linear Cox	47% (37%, 57%)	**45% (35%, 55%)**	35% (19%, 51%)
	NPI	54% (44%, 64%)	**49% (39%, 59%)**	41% (27%, 55%)
PSEP for MFP risk groups		53%	**54%**	50%
PSEP for linear Cox risk groups		46%	**45%**	47%
PSEP for NPI^{q4} risk groups		41%	**42%**	43%

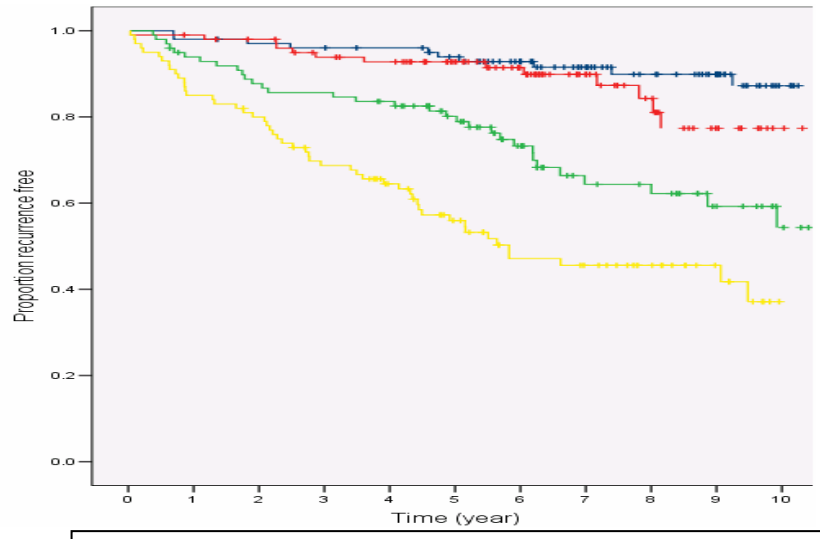
Figure 10.1: K-M survival curves for MFP (top panel) and linear Cox risk groups (bottom panel)

MFP



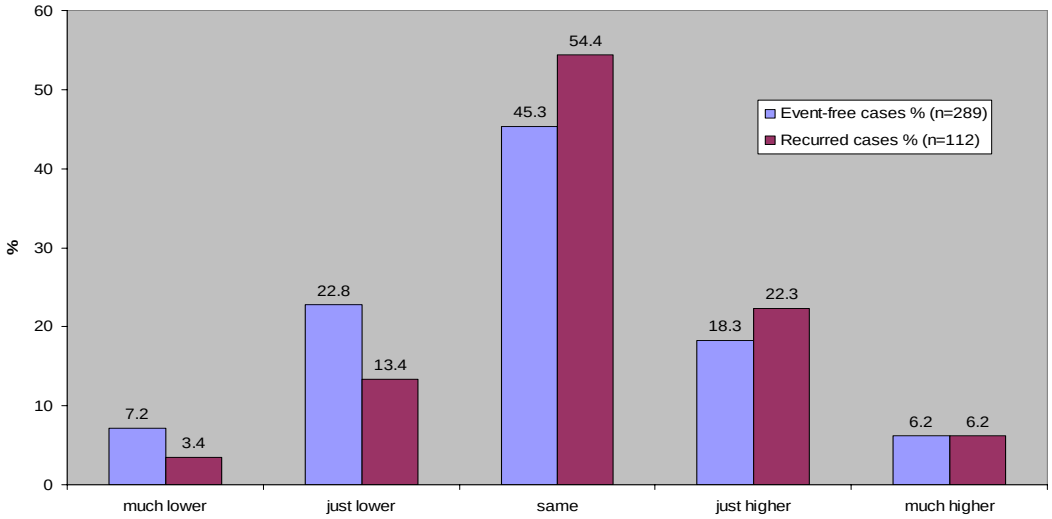
102	98	87	60	35	16
103	91	77	46	25	17
96	82	65	30	15	8
100	63	40	25	14	4

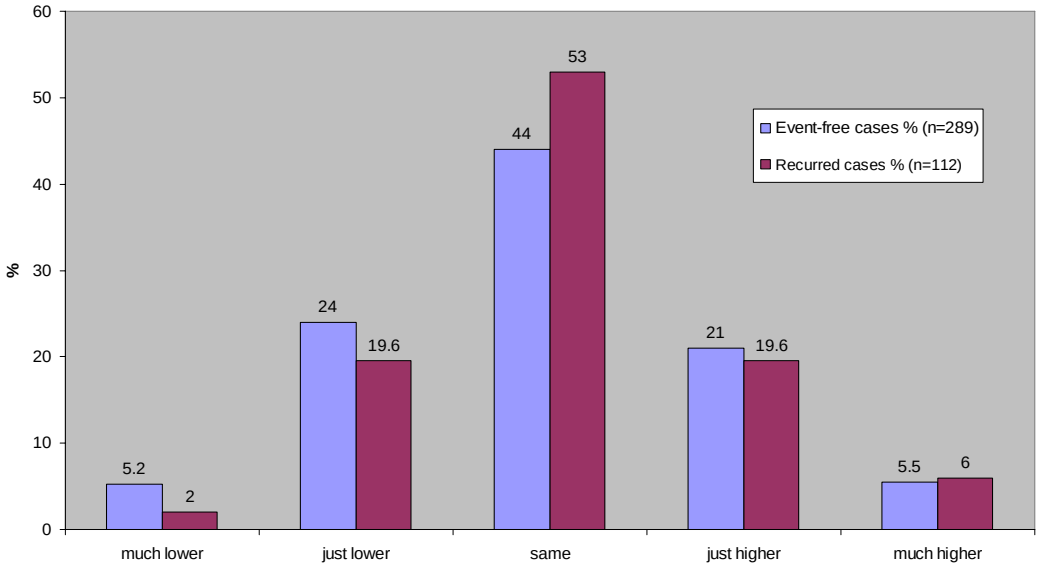
Linear Cox



101	96	88	63	42	22
101	89	74	39	16	9
99	83	65	32	19	17
100	66	42	27	12	3

Figure 10.2: Cross-classification of models separately for MFP (top panel) and linear Cox (bottom panel) risk groups against NPI^{q4}, for patients that did and did not recur





Just lower/ higher: 1 risk group shift Much lower/ higher: >1 risk group shift

10.4.3 Dichotomised biomarker models

i) Comparison between data driven Optimal split and Quartile, and pre-specified Median models

The number of variables retained in the multifactorial Optimal split, Quartile, and Median models was 6, 6, and 4 respectively. The only variables contributed to all 3 models were nodal status and tumour size (Table 10.5). Grade was retained only in Median model.

The optimal split index showed higher C-index, R-square, and goodness of fit, followed by the Quartile index (Table 10.5). On the other hand, the estimated PSEP (see section 4.4.7) for Quartile risk groups was 61%, the corresponding figures for Optimal split and Median risk groups were 55% and 42%, respectively (Table 10.6). Furthermore, estimated 7-year RFS in the lowest quartile of Quartile index was 97%, which was 4 percentage points better than that of Optimal split and 7 percentage points better than that of Median models (Table 10.6).

The results presented suggested superiority of performance for models developed in which a data-driven method was used to dichotomise the biomarkers. K-M curves corresponding to all 3 models are given in Figure 10.3. The middle 2 risk groups derived from Quartile index were very similar, but it gives the best low risk subset.

ii) Comparison of dichotomised biomarkers risk groups with NPI^{q4}

Risk group assignment by dichotomised biomarker models relative to NPI^{q4} is plotted in Figure 10.4. Similar to continuous model risk groups (Figure 10.2), performance of dichotomised biomarkers was better than NPI^{q4}.

Based on Optimal split risk groups, in comparison with NPI^{q4}, the proportion of recurred patients that moved into a more appropriate risk group was 30% and a less appropriate group was 15% (Appendix 2). Corresponding figures for non-recurred subjects were 35% and 27%, respectively. Net gain for recurred and non-recurred patients were 15 and 8 percentage points, giving NR Index of 23% (P-value=0.002).

Quartile risk groups gave comparable results with Optimal split. For subjects who had a recurrence, the risk group assignment was improved for 31% and became worse for 19% (see Appendix 2). Figures for those who did not experience the a recurrence event were 30% and 21%, respectively. Net gains were estimated at 12 and 9 percentage points, giving a NR Index of 21% (P-value=0.01).

For Median risk groups, net gain in classification improvement of recurred and non-recurred subjects were 5 and 4 percentage points, respectively (see Appendix 2) giving a NR Index of 9%, which was far from being significant (P-value=0.20).

In terms of the ability to identify low risk patients, 25% of patients in the lowest quartile of Quartile index exhibited 7-year RFS rate of 97% which was 6 percentage points higher than NPI (which had 4 equal risk groups), and 2 percentage points higher than the biological collaborators of this study expected to find.

Table 10.5: Comparison of models which deal with dichotomised data: Optimal split, Quartile, and Median models

Model	Optimal split model		Quartile model		Median model	
	HR (95% C.I.)	P-value	HR (95% C.I.)	P-value	HR (95% C.I.)	P-value
Nodal	1.89 (1.47, 2.43)	<0.001	1.96 (1.5, 2.56)	<0.001	1.89 (1.46, 2.46)	<0.001
Size (cm)	1.20 (1.10, 1.30)	<0.001	1.20 (1.10, 1.30)	<0.001	1.22 (1.10, 1.35)	<0.001
Pmtor	0.32 (0.16, 0.62)	<0.001	0.47 (0.28, 0.80)	0.006	Not screened in	
Tunel	1.97 (1.28, 3.02)	<0.001	1.93 (1.28, 2.92)	<0.001	Not screened in	
Praf338cy	1.95 (1.24, 3.07)	0.004	1.97 (1.28, 3.03)	<0.001	N.S.	
Pmapknu	2.23 (1.41, 3.54)	<0.001	Not screened in		Not screened in	
Akt2cy	N.S.		0.53 (0.31, 0.92)	0.025	N.S.	
Prhisto	N.S.		N.S.		0.62 (0.42, 0.92)	0.02
Grade	N.S.		N.S.		1.34 (1.01, 1.76)	0.04
Performance						
C-index	78.5%		76.5%		74.5%	
R-square	26%		22%		18%	
Chi-square	117		96.5		77.4	
NR Index relative to NPI^{q4}	23%		21%		9%	

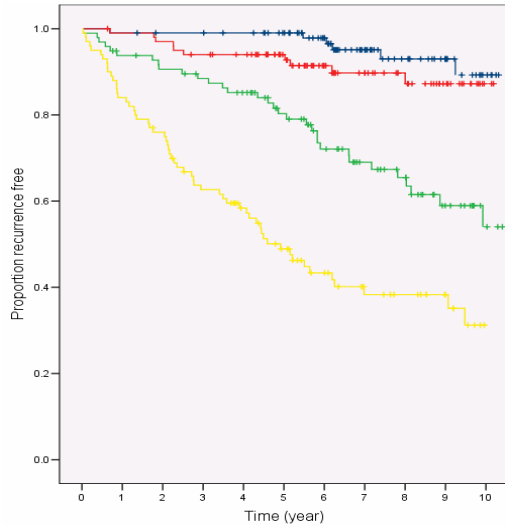
N.S. Not significant

Table 10.6: Estimated RFS rates in the lowest and highest quartiles of indices derived from dichotomised biomarker models

Risk group	Index	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	Optimal split	98% (96%- 100%)	**93% (87%- 99%)**	83% (67%- 99%)
	Quartile	97% (93%-100%)	**97% (93%-100%)**	84% (66%- 100%)
	Median	96% (92% -100%)	**90% (82%- 98%)**	79% (61%- 97%)
	NPI	95% (91%, 99%)	**91% (85%, 97%)**	84% (72%, 96%)
Highest	Optimal split	43% (33%- 53%)	**38% (28%- 48%)**	30% (16%- 44%)
	Quartile	43% (33%- 53%)	**36% (26%- 46%)**	26% (12%- 40%)
	Median	50% (40%- 60%)	**48% (38%- 58%)**	35% (21%- 49%)
	NPI	54% (44%, 64%)	**49% (39%, 59%)**	41% (27%, 55%)
PSEP for Optimal split risk groups		55%	**55%**	53%
PSEP for Quartile risk groups		54%	**61%**	58%
PSEP for Median risk groups		46%	**42%**	44%
PSEP for NPI^{q4} risk groups		41%	**42%**	43%

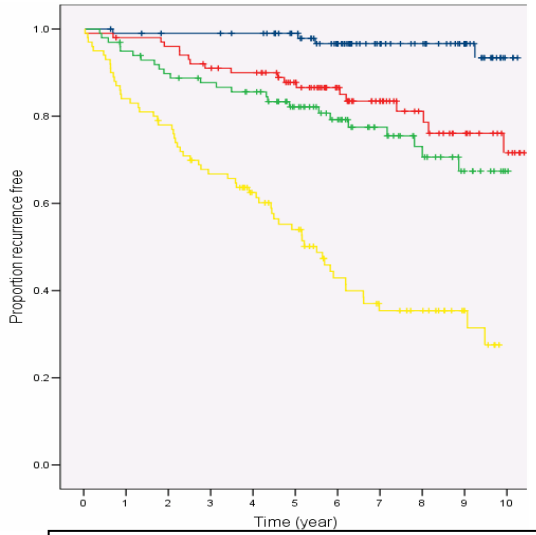
Figure 10.3: K-M curves for Optimal split (left panel), Quartile (middle panel), and Median risk groups (right panel)

Optimal split



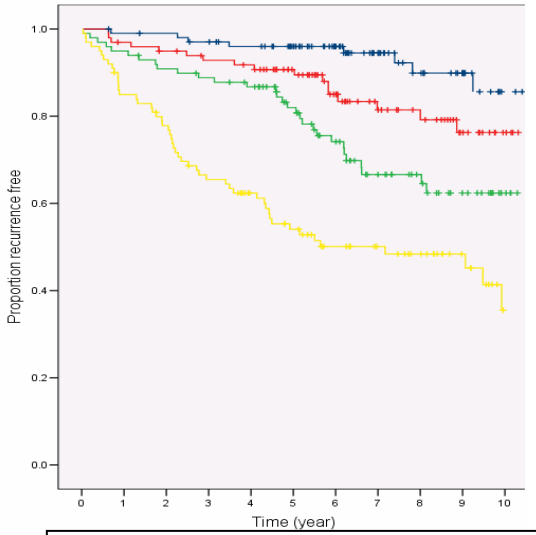
103	100	92	54	30	18
101	93	74	45	25	12
97	81	64	41	21	11
100	60	39	21	13	4

Quartile



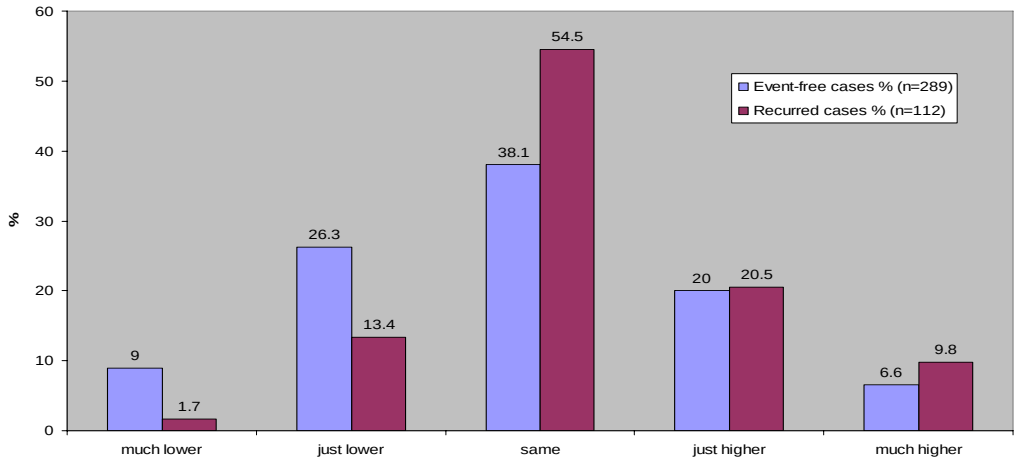
101	97	87	56	36	15
101	90	84	44	24	16
99	83	64	39	19	11
100	64	44	20	10	3

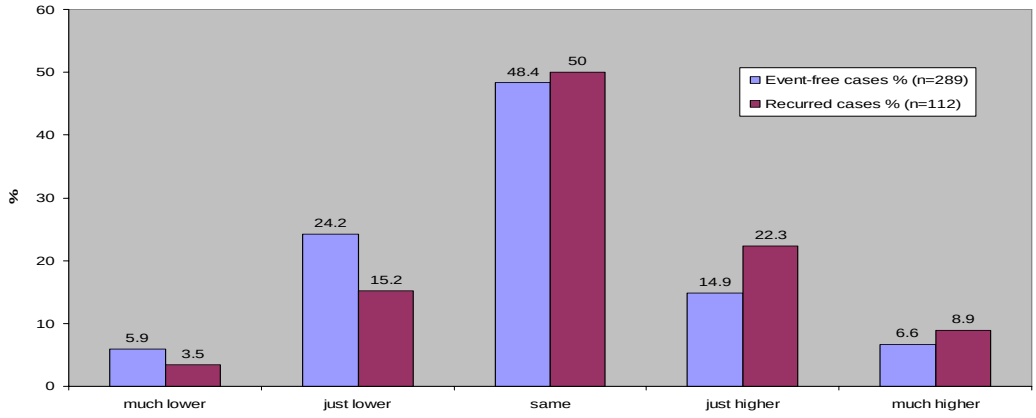
Median

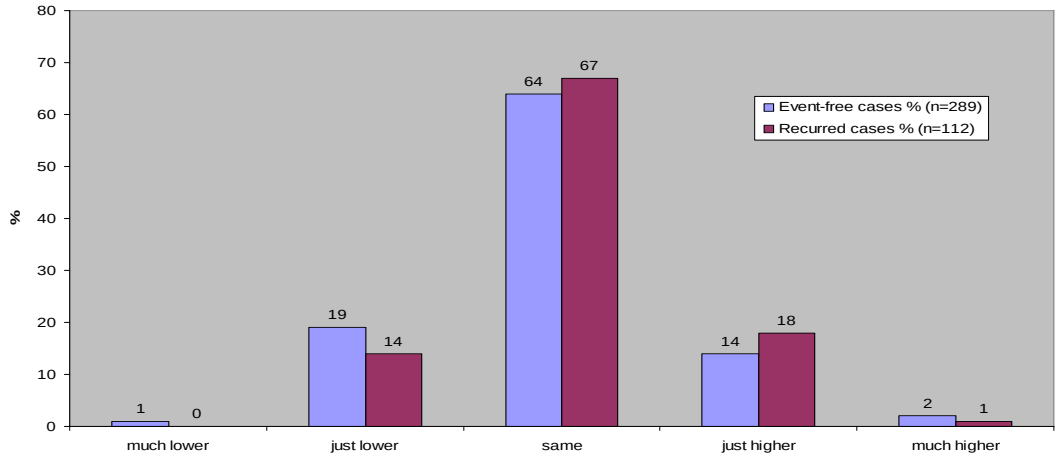


103	97	85	50	28	14
99	88	73	42	24	15
99	86	67	40	22	11
100	63	44	29	15	5

Figure 10.4: Cross-classification of models separately for Optimal split (top panel), Quartile (middle panel), and Median risk groups (bottom panel) against NPI^{q4} for patients who did and did not recur







Just lower/ higher: 1 risk group shift Much lower/ higher: >1 risk group shift

10.4.4 Comparison of models developed with the UIVS Model

Performance of the UIVS Model was superior to NPI (Chapter 8). Among models developed in this chapter, the risk groups derived from the linear Cox and Median indices, relative to NPI^{q4}, did not improved classification of patients (Table 10.7 rows 3 and 6). This also indicates their weaker performance in comparison with the UIVS Model. Therefore, no further attention was given to these two models.

i) Comparison between the MFP and UIVS Models

The MFP model, in comparison with the UIVS Model, used information on 2 fewer biomarkers. In the MFP model there were no contributions from the threshold effect of Pmapknu, or the non-ordinal effect of Akt1nu. Loss of these two biomarkers resulted in a 3.5 percentage point decrease in the C-index (Table 10.7 rows 2 and 7).

In terms of detection of low-risk patients, the number of recurrences in the lowest quartile of the MFP index was twice as large as that of UIVS (8 versus 4), but it did not affect estimated 7-year RFS rate (94% for MFP versus 95% for UIVS). This is because approximately half of those 8 recurrences happened after 7 years of follow-up, whereas all 4 recurrences observed in the bottom quartile of the UIVS index happened before the 7^{th} year. Therefore, 10-year RFS for UIVS was 95% (95% C.I.: 89%, 100%), as high as the 7-year rate, but the corresponding figure at 10 years for MFP was 80% (95% C.I.: 62%, 98%).

ii) Comparison between the Optimal split and UIVS Models

The discrimination and separation ability of indices were comparable (Table 10.7 rows 4 and 7). The 7-year RFS rate in the UIVS lowest risk group was 2 percentage points higher than Optimal split (95% versus 93%), but for both models the estimated PSEPs at 7 years were 55%. On the other hand, NR Index for Optimal split, relative to NPI^{q4}, was 5 percentage points higher than that of UIVS (Table 10.7 rows 4 and 7).

iii) Comparison between the Quartile and UIVS Models

The discrimination and separation ability, as well as goodness of fit, of the UIVS index was greater that of the Quartile index (Table 10.7 rows 5 and 6). However, the Quartile index created better diverged low and high risk patients (PSEP 61% for Quartile versus 55% for UIVS risk groups). Furthermore, estimated 7-year RFS rate for patients in the bottom quartile of the Quartile and UIVS indices were 97% and 95%, respectively.

In general, a comparison of 6 statistics (explained in section 4.4.7) showed that the results presented found that performance of MFP, Optimal split, and Quartile Models were comparable with the UIVS Model.

Table 10.7: Comparison of performance of indices and risk groups derived from continuous and dichotomised biomarker models

Row[5]	Model	C-index	R-square	Chi-square	PSEP	Classification improvement over NPI with 4 equal risk groups (NPI^{q4})								
						Recurred patients			Non-recurred patients			NR Index		
						Number (%)	Z	P-value	Number (%)	Z	P-value	Number (%)	Z	P-value
1	NPI	72%	14%	59.8										
2	MFP	75.5%	21%	88.6	54%	13 (12%)	1.81	0.08	16 (5%)	1.28	0.20	29 (17%)	2.22	0.02
3	Linear Cox	73.5%	18%	78	45%	5 (5%)	0.69	0.50	9 (3%)	0.68	0.50	14 (8%)	0.97	0.34
4	Optimal split	78.5%	26%	117	55%	17 (15%)	2.37	0.02	25 (8%)	1.87	0.06	42 (23%)	3.03	0.002
5	Quartile	76.5%	22%	96.5	61%	14 (12%)	1.87	0.06	25 (9%)	2.05	0.04	39 (21%)	2.68	0.01
6	Median	74.5%	18%	77.4	42%	5 (5%)	0.83	0.40	11 (4%)	1.09	0.28	16 (9%)	1.26	0.20
7	***UIVS***	*79%*	*28%*	*123.6*	*55%*	*16 (14%)*	*2.23*	*0.02*	*12 (4%)*	*0.90*	*0.36*	*28 (18%)*	*2.37*	*0.02*

[5] Row 1 shows performance of NPI (see Chapter 6). Rows 2 to 6 summarised performance of models developed in this chapter where a single screening method were applied to select the informative biomarkers and also form. Performance of the UIVS Model in which 3 screening methods were applied to select the informative biomarkers and form is given in row 7.

10.5 Discussion

I developed five models. The nature of the linear Cox and MFP models were similar in that they both modelled continuous data, but were different in that the form of risk function was pre-specified in the former but data-dependent in the latter. Similarly, the Optimal split, Quartile, and Median methods were applied to dichotomise the biomarkers, an analogous data driven/pre-specified contrast.

Additionally, the linear Cox and Median models were similar in that the forms of risk function/place of split were pre-specified in advance. For the MFP and Optimal split models (and to some extent Quartile model) were similar in that the form of risk function/ place of split were optimised for the data.

10.5.1 Variables contributed to the multifactorial models

Variables that contributed to the multifactorial models were: nodal status, tumour size, Pmtor, Tunel, and Praf338cy, these contributed to all models except the Median model. Grade and Prhisto were only retained in the Median model, while Krascy contributed a polynomial association only to the MFP model. Pmapknu was retained only in the Optimal split, while Akt2cy was only retained in the Quartile model.

10.5.2 Continuous data models

The Linear Cox model was the simplest model, and it performed the same as NPI. On the other hand, performance of MFP was superior to NPI (NR Index 17%, and 3 percentage points improvement in 7-year RFS of lowest risk group).

Comparison of the MFP and linear Cox regression models allowed for the contribution of polynomial effects on improvement of performance of the multifactorial models to be addressed. It has been recommended that to make the most use of information, data should be explored to investigate whether transformation of variables can reveal more information and whether it improves the fit [Knorr KL et al., 1992]. Considering non-linear patterns is very important, if these kinds of relationships are ignored then an important variable might show a non-significant linear effect [Hastie T et al., 1992]. Results presented indicated that MFP improved the performance of the model, this was not against my expectation since MFP detects transformations which give the best fit to the data.

10.5.3 Dichotomised data models

In Median model, the place of split was blind to the data, and performance of the model was similar to NPI. On the other hand, once place of split was optimised (Optimal split model), a noticeable improvement in the C-index and NR Index was seen. This might not be surprising since in the Optimal split model, each biomarker was dichotomised so as to separate low and high risk patients as much as possible.

However, interestingly the Quartile index worked much better than NPI in terms of risk group assignment, and the ability to distinguish low and high risk patients. The number of recurrences observed in the lowest quartile of the Quartile index was 5, while the number of relapsed patients in the lowest risk group of NPI^{q4} and NPI^{std3} were 10 and 15, respectively.

10.5.4 Comparison of continuous versus dichotomised models

Royston *et al.*, in analysing 3 continuous and 2 binary variables criticised dichotomisation of continuous data [Royston P et al., 2006]. For all analyses done by Royston *et al.*, 17 percent of subjects with missing data were excluded. Applying the MFP procedure, the final model comprised 2 continuous (one with linear form and one with polynomial form) and 2 binary variables, the third continuous variable did not remain in the MFP model. Focusing on these 4 significant variables in MFP model, 2 more models were developed that dichotomised continuous predictors at optimal split and at median. The MFP model, in which continuous variables were treated in continuous form, gave higher predictive and discrimination ability and larger model chi-square [Royston P et al., 2006].

The variables that contributed to Roystons' study were the same. However, in my analyses, the variables that contributed significantly to the multifactorial models were not the same. Comparing performance of all 5 models developed, main findings were as follows:

In terms of goodness of fit, discrimination and separation ability, and risk group assignment relative to NPI, the indices and risk groups derived from dichotomised data-driven models (Table 10.7 rows 4 and 5) showed higher performance than either data-driven (row 2) or pre-specified (row 3) continuous indices. Results of linear Cox and Median indices, in which pre-specified form/ split was applied, were the same and inferior to other models (rows 3 and 6). Better performance of the Optimal split model (row 4) might be explained by the fact that this model dichotomised each biomarker to separate low and high risk patients as much as possible, and therefore results might be overly optimistic. Furthermore, by chance, the upper quartile of biomarkers that contributed to the Quartile model was similar to that of the optimal split. That is why the Quartile model gives estimates close to the optimal split, and marginally better than models using continuous biomarkers.

In terms of ability to identify low risk patients, 25% of patients that fell into the lowest quartile of the Quartile index exhibited the best 7-year survival (97%). A subset of low risk patients with only 5 recurrences was detected. Performance of lowest risk group derived from MFP and Optimal split indices were comparable.

10.5.5 Role of Akt2cy in detection of low risk patients

Lowest risk groups derived from the Quartile index exhibited the best survival. The main difference between this and other indices was the contribution of Akt2cy in the Quartile model. To check the contribution of this biomarker to the detection of good prognosis patients, I refit the Quartile model without use of Akt2cy. Exclusion of Akt2cy strongly decreased 7-year RFS in the lowest risk group (91% versus 97%).

The number of recurrences falling into lowest risk group in the absence and presence of Akt2cy was 11 and 5, respectively, indicating a substantial role for this biomarker in identification of low-risk patients. Furthermore, in the absence of Akt2cy, NR Index reduced from 21% to 15%, which was still of marginal significance (P-value=0.04).

10.5.6 Comparison of biomarkers and the UIVS Models

Results suggest that better performance of the UIVS model, in comparison with NPI, was not due to application of comprehensive and complicated screening procedures. The MFP, Optimal split, and Quartile models also produced results comparable with the UIVS, however results of the Optimal split model are over-optimistic due to an extensive cut point search. In addition, the Quartile model was not an extreme data-dependant model since only 3 cut points were tested to select the best split. The UIVS index showed higher discrimination and predictive ability, but risk groups derived from Quartile index were better diverged.

Screening procedures depend on the aim of the study. When the aim is to generate new biological questions, in-depth analysis is more appropriate. On the other hand, when the aim is simply outcome prediction, then application of MFP might be enough. In this data set, dichotomisation of biomarkers at upper quartile also led to a model with good performance. Comparison of results of in-depth with simple screening methods can reveal the value and necessity of in-depth analysis.

10.6 Chapter summary

- With biomarkers in continuous form, the MFP model showed better performance than the linear Cox model. The main difference between the models was the inclusion of Krascy with polynomial effect in the MFP.

- When I dichotomised the data, the Optimal split index had higher discrimination and separation ability than Median index. This was expected since Optimal split model used the split which best separated subjects into 2 risk groups.

- The Quartile model gave results similar to the Optimal split model. It appears that, by chance, for a number of biomarkers the Q3 values were very close to optimal split.

- When modelling skewed biomarkers, data-driven dichotomised models worked better than continuous models. However, results are prone to be overoptimistic due to the nature of data-dependent minimum P-value method used in development of optimal split model.

- Good performance of the UIVS Model was not simply due to a complex screening phase. MFP and Quartile models gave results comparable to the UIVS Model.

Chapter 11 OVERALL DISCUSSION

11.1 Introduction

Methodological issues presented in this thesis are applicable to a range of data set types, and are not specifically designed for breast cancer or for follow-up data sets. Estimation of the optimum form, imputation of missing data, utilisation of appropriate methods to deal with many variables, and assessment of internal validity are practical challenges in a wide range of regression settings (continuous, binary and time-to-event outcomes).

A rich array of biomarkers with potential relevance to cancer progression was available. The current data set was a typical data set in the extent of biomarkers and follow-up it has, and follow-up is continuing. In addition, the methodological

developments presented in this thesis also allow modelling gains through detection of optimum form of association for skewed biomarkers, and provides powerful multiple imputation methods to salvage as much predictive information as possible from subjects with missing data.

Professor John Bartlett had 2 main questions. He wanted to know, out of 72 biomarkers, which had the potential to predict Recurrence Free Survival (RFS). His second question was that whether a combination of informative biomarkers and clinical variables (nodal status, grade, and tumour size), which are used in development of Nottingham Prognostic Index (NPI), improved the ability of NPI to detect patients with very low risk of recurrence.

There is very scant information about the role of biological aspects in tumour progression. In the exploratory phase of an explanatory prognostic study, when the aim is to describe the associations as best as possible and to generate questions about biology of disease, data should be explored to extract as much information as possible [Hayden JA et al., 2008]. That is why I applied a range of methods to estimate form of association and to develop biomarker models. In Royston words 'Collecting data is expensive whereas fitting models is cheap. There is room in science for trying several approaches with a given data set and reviewing the results critically' [Royston P et al., 2000].

The original statistical contribution of this thesis was to combine the MICE and bootstrap procedures in the presence of non-linear effects. Heymans *et al.* combined

the MICE and bootstrap but assumed that effects are linear [Heymans MW et al., 2007]. Royston et al. emphasized the use of FP to explore potential polynomial effects followed by bootstrapping to avoid unstable results [Royston P and Sauerbrei W, 2003; Sauerbrei W and Royston P, 2007]. However, no previous study dealt with all 3 of these issues in one modelling process.

The advantages and disadvantages of the processes developed in each chapter were discussed previously. This chapter is a comparative discussion of processes and other general issues, such as importance of investigation of stability of effects, and also ways to combine estimates across multiply imputed data sets. Additionally, some topics for future research are proposed.

11.2 Statistical issues

Two main strategies were developed. In the Univariately Informative Variable Selection (UIVS) approach, an in-depth screening was applied to select potentially informative biomarkers and form of association. In the Biologically Guided Variable Selection (BGVS) approach, model building was guided by biological knowledge where substantive sets of biomarkers were created. Performance of the BGVS Model was superior to the UIVS Model, highlighting the importance of use of external information in the process of model development.

The process applied to develop the UIVS Model was explored in detail in Chapters 9 (by applying all variables to the model and by replacing missing data with median

values) and 10 (by simplifying the screening process used to select informative biomarkers and form). However, no further exploration was performed for the BGVS Model. This is because one of the components of the BGVS process was to use biological expertise which might not be widely available. The BGVS Model was only developed to explore the value of close co-operation between statisticians and biologists in the process of model development.

Results presented in Chapter 9 indicated that when informative biomarkers were submitted to the multifactorial model, substitution of missing data by median was a good approximation for the MICE method, probably due to the low missing value rate. However, it will not be possible to verify this fact until a range of methods is applied.

Results presented in Chapter 10 suggested that use of Fractional Polynomial (FP) to select informative biomarkers and form, followed by application of Multivariable Fractional Model (MFP) to develop the multifactorial model, provided results comparable to the UIVS Model. Furthermore, in this data set, the upper quartile was a reasonable split to dichotomise biomarkers, providing a model which performed similarly to the UIVS Model. Interestingly, the Quartile model was able to detect a low risk group with lower risk of recurrence at 7 years than that of the UIVS. When I explored the Quartile model, I saw that Akt2cy had a substantial role in identification of low risk patients. This biomarker was not selected in any of the other models, except the BGVS. Therefore, when the aim is to understand underlying biological mechanisms, I recommend developing a complex model allowing for all possible

forms of association, and to compare its performance with simplified models. This increases understanding of risk factors that govern the disease course, which will not be obtained unless a range of models are developed. These all provide new questions to be tested in independent data sets.

11.2.1 Investigation of stability of transformations

The importance of investigation of stability of non-linear effects is highlighted in section 8.5.4. A simple approach would be to apply the transformations found in the screening phase to the imputed data sets, and to use the transformed variables in development of a multifactorial model. As an example, this was done in two papers published in STATA journal [Royston P, 2004; Royston P, 2005], I feel this is because in the STATA bulletin emphasis is on the practical usage of special packages rather than methodological issues. Royston and Sauerbrei recommended the use of bootstrap procedure to extract more information from the data [Royston P and Sauerbrei W, 2003; Sauerbrei W and Royston P, 2007].

As explained in section 8.5.4, when I applied the transformations and a multifactorial Cox model, all of the 5 non-linear effects were retained in the model. However, bootstrap stability check showed that two of them were unreliable. In my experience, use of transformed data in a multifactorial model, without stability checking, has three main disadvantages:

1. Variables with unreliable transformation might contribute to the multifactorial model increasing risk of overfitting and model instability

2. A variable with univariate polynomial effect might show different behaviour in multifactorial analysis

3. If the case missing values rate is high, the form found in a univariate analysis (when missing data for variable being tested is excluded), might not express effect of variable or might not be optimum even in univariate analysis of imputed data sets (after imputation of missing data)

11.2.2 Aggregation of forms and coefficients

Multiple imputation of missing data is frequently used in the literature. However, when working with multiple data sets, the methods to tackle many practical issues are still open to discussion [Royston P and Sauerbrei W, 2008]. One of the most important problems is that, when working with multiply imputed data sets, the variables retained in each model and the appropriate form of risk function might not be the same across all data sets. Therefore, it is necessary to make decisions about the appropriate form and then to combine results from each imputation so as to give a single model [Royston P and Sauerbrei W, 2008]. My work is one of the first movements in that direction.

i) Only linear forms exist

When all forms are linear, only aggregation of parameters (regression coefficients and standard errors) and risk scores is required. There are 3 main ways to aggregate results across multiply imputed data sets.

The first solution is to calculate a data set specific risk score for each of the 10 imputed data sets. The average of 10 risk scores can be used as the final index. In this case, it is not possible to calculate single HR's and to specify a prognostic formula.

The second method is to aggregate parameter estimates via Rubin's rule (section 4.4.4 part i). These aggregated estimates then can be used to calculate Hazard Ratios (HR), and then applied to each imputed data set to calculate a risk score. Average of calculated risk scores can be used as the final prognostic formula.

The third solution is to aggregate parameter estimates via Rubin's formula (section 4.4.4 part i), to calculate HR's, but to use average of data set specific risk scores as the final index. Therefore, it would not be possible to give a prognostic formula.

In my opinion, since aggregated parameter estimates do not provide the best fit to any of the imputed data sets, application of this method might result in the risk groups not having the best possible diverged curves. However, it remains possible to specify a prognostic formula. On the other hand, use of the average of dataset specific risk scores might improve risk stratification, but no single prognostic

formula can be given. This issue needs further exploration in future studies (see section 11.4.2).

ii) Non-linear forms exist

The process of aggregation of forms and parameter estimates will be more complex when non-linear forms exist (such as this project). This is because the form of association might not be the same in all bootstrap or imputed data sets. Therefore, aggregation of estimated coefficients for a single variable with different forms across samples is not possible.

The algorithm I devised in the process of development of the UIVS Model was as follows. To aggregate the forms, I applied forms repeated in the majority of bootstrap samples (> 50%) to each of the 10 imputed data sets. I then estimated parameters and aggregated them across the 10 imputed data sets applying Rubin's formal. Aggregated estimates were used to calculate HR's. However, to present Kaplan-Meier survival curves, I used dataset specific regression coefficients so as to maximise the separation between risk groups.

The reason I used dataset specific regression coefficients, instead of aggregated coefficients, was that one of the clinical questions Professor John Bartlett asked was whether a combination of biomarkers and clinical variables improves the ability of NPI in terms of identification of low risk patients. Therefore, it was very important to apply coefficients which increase the chance of finding such a low risk group. This is because if the risk groups derived do not show improvement over NPI in this training sample, then there would be no point to assess the external validity of the model in

an independent sample. I feel that use of data specific coefficients is superior to aggregated ones in terms of risk stratification.

Some alternative solutions with advantages and disadvantages are given below. If one applies a dataset specific form of association and dataset specific regression coefficients to each of the imputed or bootstrapped samples, then the variables and forms contributing to single risk scores are very likely to be different. In this case, although the average of risk scores can be used as a final index to categorise patients into risk groups, Hazard Ratios (HR) cannot be calculated and no single prognostic formula can be specified.

On the other hand, to be able to calculate HR, and also to specify a single prognostic formula, one can apply the form repeated in the majority (> 50%) of the samples, plus the aggregated estimates across samples. I devised this approach to check whether application of aggregated coefficients affects separation of risk groups. Interestingly, almost no difference in estimated RFS rates in the lowest and highest risk groups was seen (data not shown). This can be explained by the fact that the missing rate was very low and therefore the dataset estimated coefficients were very similar to aggregated estimates. However, this might not be the case when missing rate is high.

In my opinion, it is not simple to advise the use of any one of the approaches explained above for all future research and data sets. This is because statisticians would like to calculate CI's and precision of parameter estimates, but clinicians are

more interested in a single prognostic formula. Decisions can be made by balancing between statistical and clinical purposes of the study.

In general, following from the arguments given above, the application of a data specific form might not appeal to either statisticians or a clinical audience. Application of form repeated in the majority of samples and dataset specific coefficients partly solve this problem, since calculation of HR is possible but no single prognostic formula can be given. This can be tackled by using forms repeated in the majority of samples and aggregated estimates. However, there is no guarantee that the use of aggregated estimates will produce the best separation of risk groups.

11.3 Clinical issues

Using retrospective statistical modelling of extensive molecular pathways, I developed a series of prognostic models which stratified early breast cancers treated with tamoxifen by risk of recurrence.

Individual risk scores were calculated using a simple panel of biomarkers (6 for the UIVS and 8 for BGVS Models, in addition to tumour size and nodal status), this has the potential to provide a cost effective and readily applicable platform for future diagnostic applications. When patients were stratified by risk into 4 quartiles, marked differences in group relapse rates were observed.

NPI is the recognised tool for risk prediction used in the UK. These findings demonstrate considerable potential for improved prognostic modelling by incorporation of biological variables into risk prediction:

i) Ability to detect low risk patients

Although at 5-years follow up, the Recurrence Free Rate (RFS) was similar for the lowest risk groups by NPI and the biomarker models, longer-term RFS appears better predicted. Among patients in the lowest risk quartile of the UIVS index (i.e. with the lowest risk scores) the estimated 7-year RFS rate was 95%, whereas in the highest risk group (top quartile) it was only 40%. Even when I applied cut offs to create risk groups similar to standard NPI with 3 risk groups NPI (NPI^{std3}), 133 were fell into the lowest risk group giving 7-year RFS of 95% (95% C.I.: 91%, 99%).

Results were even better when family risk scores were modelled (the BGVS Model). Actuarial 7-year RFS for patients in the bottom and top risk quartiles was 98% (95%C.I.: 96%, 100%) and 40% (95%C.I.: 30%, 50%) respectively. When risk groups similar to NPI^{std3} were created, estimated 7 and 10-year RFS in lowest risk groups was 96% (95% C.I.: 92%, 100%) at both time points. This indicated that one-third of subjects had sufficiently low risk of disease recurrence. This offers a clinically reassuring recurrence free rate in the lowest risk group and has clinical potential, since it appears such patients might be spared harsh treatments, without undue risk of recurrence. However, the very small number of events and a limited number of patients with follow-up exceeding 8 years in this cohort means that at this stage interpretation of results past 7 years should be done with caution.

ii) Classification improvement

Biomarker Models (UIVS and BGVS) demonstrated significantly better risk group classification performance than NPI. There was overall a greater number of improved risk group classifications than less appropriate ones, this was more apparent in the subgroup who went on to experience recurrence (percentage points difference of 14 for the UIVS and 21 for the BGVS). Corresponding NR Indices were 18% and 32% respectively.

iii) Treatment selection

Endocrine therapy, using either tamoxifen or aromatase inhibitors, remains the most successful approach to the treatment of early breast cancer. However, many women do not require even endocrine therapy, or might derive minimal additional benefit over tamoxifen treatment if treated with aromatase inhibitors and/or chemotherapy [Abe O et al., 2005].

Powerful predictive biomarker tools have been proposed that have potential for future application in the selection of patients for conservative versus aggressive adjuvant treatment. Whilst low risk patients selected by biomarker models could potentially avoid systemic treatment, higher risk patients might require additional treatment, including chemotherapy or other adjuvant treatment options.

NPI is a parsimonious model based on 3 variables. Parsimonious models, based on a small number of variables, are fairly stable and can easily be applied in practice [Van Houwelingen HC, 2000; Laupacis A et al., 1997]. Although this model is very

useful, it was devised using a fairly limited range of risk factors on a cohort with fairly short follow-up time. However, in some ways it is surprising that it has not already been superseded.

The approaches developed have been to use markers of key molecular pathways of tamoxifen resistance to seek to identify a panel which can select patients who may either derive sufficient benefit from treatment with tamoxifen alone, or for whom withdrawal of even this moderately toxic adjuvant therapy might pose minimal risk.

The models developed provide a significant potential improvement over conventional prognostic model (NPI) and needs to be validated in a larger clinical trial cohort. However, further validation of the statistical approaches undertaken is also required. With larger data sets and longer follow-up this modelling approach has the potential to enhance understanding of the interplay of biological characteristics, treatment and cancer recurrence. If biomarker modelling of other breast cancer cohorts can be undertaken, the statistical modelling method illustrated here promises to aid understanding and prevention of breast cancer progression.

11.4 Studies for future

11.4.1 Assessment of external validity of the model

I developed reliable approaches, but using a fairly small data set with limited follow-up. The results reported here can be used to generate hypotheses about the mechanisms that govern breast cancer. However, what would be interesting, both to understanding the biology of progression and recurrence, and to improving risk prediction, would be to validate the models and to assess the generalisability of results in an independent larger data set with more extensive follow-up. Results presented in this thesis are tentative until a future validation set is available.

Generalisability or external validity means how well the model works in case of future patients [Justice AC et al., 1999]. Altman *et al.* noted that a useful model is the one which works in practice and not the one with many zeros in the corresponding P-values in the multifactorial model [Altman DG and Royston P, 2000].

Assessment of external validity, before implementation of the model in practice, is of crucial importance. Hence, further validation of the models presented is planned in a large patient cohort. Further markers might be included, as appropriate, to extend the applicability of the model. An important issue to be considered in a validation study is the extent to which derivation and validation samples are similar. An important reason that a prognostic model does not work in independent samples is heterogeneity between the derivation and validation populations [Bleeker SE et al., 2003].

The cohort used in my project had of a fairly narrow range of patient presentations (ER+ tamoxifen treated). This can be a strength, since a homogeneous set of patients was analysed, and potentially a narrower range of mechanisms and predictive variables needs to be included; and therefore the model has the potential to provide insights to cancer biologists regarding the mechanism of cancer progression in this patient group. Furthermore, it is likely that as medical knowledge grows, separate prognostic models will be sought for the main different patient subgroups, such as this cohort. In contrast, the fairly narrow cohort can also be a weakness, in that the prognostic performance might not generalise to other presentations.

11.4.2 Aggregation of forms and parameter estimates

Behaviours of aggregation methods discussed in section 11.2.2 needs to be explored to investigate their weaknesses and strengths. In my data set, application of forms repeated in the majority of samples and dataset specific coefficients gave results comparable to that of aggregated coefficients; the separation of risk groups was probably due to a low missing rate. The current data set can be used to omit some of the data so as to increase the missing rate and check properties of these aggregation schemes.

11.4.3 Risk classification

Although NR Index is sensitive to benefit of addition of new risk factors and could be communicated simply with a clinical audience, it can only be calculated for symmetric tables. As an example, NPI classifies patients into three groups but

biomarker models into four. Therefore the frequency distribution table is a 4 x 3 matrix and it might not be straightforward to calculate an NR Index.

Another limitation is that, the NR Index gives equal weighting to false positive and false negative classifications [Vickers AJ et al., 2009]. It has been commented that, 'a marker that results in 50 fewer biopsies of men without cancer but also 30 fewer biopsies of men with cancer will give a positive classification improvement. However, few clinicians would be prepared to miss 30 cancers in order to avoid 50 unnecessary biopsies'.

11.4.4 Comparison of screening methods for skewed variables

As mentioned in Chapter 3, alternative statistical techniques were performed to capture different simulated risk functions [Hollander N and Schumacher M, 2006]. In general, the Fractional Polynomial (FP) method performed the best. Univariate results presented in this thesis suggested that the same is true in the case of highly skewed data. However, there is scope to investigate this issue through simulation analysis and multifactorial models.

11.5 Recommendations

The aim of this thesis was not to compare and apply all possible statistical methods. I applied methods that are frequently used in the literate and then combined them to develop pragmatic strategies, which are useful for clinical purposes. In my experience, the following steps should be followed in development of a prognostic model when many continuous variables, including missing data, are available:

1. Application of univariate screening methods, to select a reduced set of variables, is the standard method. However, application of methods such as Cox PH model might lead to a loss of variables that have predictive ability.

2. Assessment of the appropriate form of association, by application of methods such as FP, enhances the understanding of biology of disease.

3. In the development of the multifactorial models, assessment of the stability of associations and exclusion of variables with unreliable association, via bootstrapping, improves model stability.

4. If external biological knowledge is available, close collaboration of statisticians and biologists should lead to models with better performance and higher biological interpretability.

5. To deal with missing data, the MICE method should be applied. Comparison of results with other simpler methods enriches the body of the literature and enhances the understanding of the value of the statistical methods.

6. Using different methods assess the form of association can generate new biological questions to be tested in independent samples.

7. When the form of association for a particular variable is different across imputed samples, a decision about how to combine the results across samples should be made by balancing between clinical and statistical aims of the study.

Reference List

Abdolell M, LeBlanc M, Stephens D, Harrison RV (2002) Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Stat Med* **21**: 3395-3409

Abe O, Abe R, Enomoto K, Kikuchi K, Koyama H, Masuda H, Nomura Y, Sakai K, Sugimachi K, Tominaga T, Uchino J, Yoshida M (2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365**: 1687-1717

Altman DG (1998) Suboptimal analysis using 'optimal' cutpoints. *Br J Cancer* **78**: 556-557

Altman DG, Andersen PK (1989) Bootstrap investigation of the stability of a Cox regression model. *Stat Med* **8**: 771-783

Altman DG, Bland JM. (2007) Missing data. *BMJ* **334**: 424

Altman DG, Lausen B, Sauerbrei W, Schumacher M (1994) Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* **86**: 829-835

Altman DG, Lyman GH (1998) Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat* **52**: 289-303

Altman DG, Royston P (2000) What do we mean by validating a prognostic model? *Stat Med* **19**: 453-473

Altman DG, Royston P (2006) The cost of dichotomising continuous variables. *BMJ* **332**: 1080

Ambler G and Benner, A. (2008) mfp: Multivariable Fractional Polynomials URL:.http://stat.ethz.ch/CRAN/

Ambler G, Omar RZ, Royston P (2007) A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat Methods Med Res* **16**: 277-298

Ambler G, Royston P (2001) Fractional polynomial model selection procedures: investigation of type one error rate. *Journal of statistical computation and simulation* **69**: 89-108

Austin PC, Tu JV (2004a) Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol* **57**: 1138-1146

Austin PC, Tu JV (2004b) Bootstrap methods for developing predictive models in cardiovascular research. *American Statisticians* **58**: 131-137

Balslev I, Axelsson CK, Zedeler K, Rasmussen BB, Carstensen B, Mouridsen HT (1994) The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). *Breast Cancer Res Treat* **32**: 281-290

Banerjee M, George J, Song EY, Roy A, Hryniuk W (2004) Tree-based model for breast cancer prognostication. *J Clin Oncol* **22**: 2567-2575

Barzi F, Woodward M (2004) Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *Am J Epidemiol* **160**: 34-45

Blamey RW, Davies CJ, Elston CW, Johnson J, Haybittle JL, Maynard PV (1979) Prognostic factors in breast cancer -- the formation of a prognostic index. *Clin Oncol* **5**: 227-236

Blamey RW, Ellis IO, Pinder SE, Lee AH, Macmillan RD, Morgan DA, Robertson JF, Mitchell MJ, Ball GR, Haybittle JL, Elston CW (2007a) Survival of invasive breast cancer according to the Nottingham Prognostic Index in cases diagnosed in 1990-1999. *Eur J Cancer* **43**: 1548-1555

Blamey RW, Pinder SE, Ball GR, Ellis IO, Elston CW, Mitchell MJ, Haybittle JL (2007b) Reading the prognosis of the individual with breast cancer. *Eur J Cancer* **43**: 1545-1547

Bleeker SE, Moll HA, Steyerberg EW, Donders AR, rksen-Lubsen G, Grobbee DE, Moons KG (2003) External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* **56**: 826-832

Bloom HJ, Richardson WW (1957) Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *Br J Cancer* **11**: 359-377

Bradburn MJ, Clark TG, Love SB, Altman DG (2003) Survival analysis Part III: multivariate data analysis -- choosing a model and assessing its adequacy and fit. *Br J Cancer* **89**: 605-611

Breastcancer.org (2008a) DCIS - Ductal Carcinoma In Situ. URL: http://www.breastcancer.org/symptoms/dcis/

Breastcancer.org (2008b) Local Treatments for IDC: Surgery and Radiation Therapy. URL: http://www.breastcancer.org/symptoms/types/idc/treatment/local.jsp

Breastcancer.org (2008c) Symptoms & Diagnosis. URL: http://www.breastcancer.org/symptoms/

Breastcancer.org (2008d) Systemic Treatments for IDC: Chemotherapy, Hormonal Therapy, Targeted Therapies. URL: http://www.breastcancer.org/symptoms/types/idc/treatment/systemic.jsp

Breastcancer.org (2008e) Treatment for IDC. URL: http://www.breastcancer.org/symptoms/types/idc/treatment/

Breastcancer.org (2008f) What is breast cancer. URL: http://www.breastcancer.org/symptoms/understand_bc/what_is_bc.jsp#Stages

Brown J, Jones M, Benson EA (1993) Comment on the Nottingham Prognostic Index. *Breast Cancer Res Treat* **25**: 283

Burton A, Altman DG (2004) Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* **91**: 4-8

Callagy GM, Pharoah PD, Pinder SE, Hsu FD, Nielsen TO, Ragaz J, Ellis IO, Huntsman D, Caldas C (2006) Bcl-2 is a prognostic marker in breast cancer independently of the Nottingham Prognostic Index. *Clin Cancer Res* **12**: 2468-2475

Cancer Research UK (2006) Breast Cancer: UK breast cancer statistics. URL: http://info.cancerresearchuk.org/cancerstats/types/breast/

Cancer Research UK (2007) UK cancer incidence statistics. URL: http://info.cancerresearchuk.org/cancerstats/incidence/?a=5441

Cancer Research UK (2008) Breast Cancer Key Facts. URL: http://info.cancerresearchuk.org/cancerstats/types/breast/

Cannings E, Kirkegaard T, Tovey SM, Dunne B, Cooke TG, Bartlett JM (2007) Bad expression predicts outcome in patients treated with tamoxifen. *Breast Cancer Res Treat* **102**: 173-179

Carroll KJ (2003) On the use and utility of the Weibull model in the analysis of survival data. *Control Clin Trials* **24**: 682-701

Chen W, Foran DJ (2006) Advances in cancer tissue microarray technology: Towards improved understanding and diagnostics. *Anal Chim Acta* **564**: 74-81

Ciampi A, Lawless JF, McKinney SM, Singhal K (1988) Regression and recursive partition strategies in the analysis of medical survival data. *J Clin Epidemiol* **41**: 737-748

Clark TG, Altman DG (2003) Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol* **56**: 28-37

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis part IV: further concepts and methods in survival analysis. *Br J Cancer* **89**: 781-786

Collett D (2003) Modelling Survival Data in Medical Research. Chapman and Hall/CRC: Florida

Colomer R, Beltran M, Dorcas J, Cortes-Funes H, Hornedo J, Valentin V, Vargas C, Mendiola C, Ciruelos E (2005) It is not time to stop progesterone receptor testing in breast cancer. *J Clin Oncol* **23**: 3868-3869

Concato J, Feinstein AR, Holford TR (1993) The risk of determining risk with multivariable models. *Ann Intern Med* **118**: 201-210

Cook EF, Goldman L (1984) Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. *J Chronic Dis* **37**: 721-731

Coradini D, Boracchi P, Daidone MG, Pellizzaro C, Miodini P, Ammatuna M, Tomasic G, Biganzoli E (2001) Contribution of vascular endothelial growth factor to the Nottingham prognostic index in node-negative breast cancer. *Br J Cancer* **85**: 795-797

Cox DR (1972) Regression models and life tables. *Journal of royal statistical society* **34**: 187-220

Croy CD, Novins DK (2005) Methods for addressing missing data in psychiatric and developmental research. *J Am Acad Child Adolesc Psychiatry* **44**: 1230-1240

D'Agostino RB, Belanger AJ, Markson EW, Kelly-Hayes M, Wolf PA (1995) Development of health risk appraisal functions in the presence of multiple indicators: the Framingham Study nursing home institutionalization model. *Stat Med* **14**: 1757-1770

D'Eredita' G, Giardina C, Martellotta M, Natale T, Ferrarese F (2001) Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution. *Eur J Cancer* **37**: 591-596

Dannegger F (2000) Tree stability diagnostics and some remedies for instability. *Stat Med* **19**: 475-491

Derksen S, Keselman J (1992) Backward, forward, and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British journal of mathematical and statistical psychology* **45**: 265-282

Donders AR, van der Heijden GJ, Stijnen T, Moons KG (2006) Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* **59**: 1087-1091

Donner A (1982) The relative effectiveness of procedures commomly used in multiple regression analysis for dealing with missing values. *American Statisticians* **36**: 378-381

Eden P, Ritz C, Rose C, Ferno M, Peterson C (2004) "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer* **40**: 1837-1841

Fairclough DL (2004) Patient reported outcomes as endpoints in medical research. *Stat Methods Med Res* **13**: 115-138

Faraggi D, Simon R (1996) A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Stat Med* **15**: 2203-2213

Faratian D, Bartlett JM (2008) Predictive markers in breast cancer--the future. *Histopathology* **52**: 91-98

Ferrandina G, Scambia G, Bardelli F, Benedetti PP, Mancuso S, Messori A (1997) Relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients: a meta-analysis. *Br J Cancer* **76**: 661-666

Fisher B, Costantino JP, Wickerham DL, Cecchini RS, Cronin WM, Robidoux A, Bevers TB, Kavanah MT, Atkins JN, Margolese RG, Runowicz CD, James JM, Ford LG, Wolmark N (2005) Tamoxifen for the prevention of breast cancer: current status of the National Surgical Adjuvant Breast and Bowel Project P-1 study. *J Natl Cancer Inst* **97**: 1652-1662

Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Daly M, Wieand S, Tan-Chiu E, Ford L, Wolmark N (1998) Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst* **90**: 1371-1388

Galea MH, Blamey RW, Elston CE, Ellis IO (1992) The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* **22**: 207-219

Goodwin PJ, Ennis M, Pritchard KI, Koo J, Trudeau ME, Hood N (2003) Diet and breast cancer: evidence that extremes in diet are associated with poor survival. *J Clin Oncol* **21**: 2500-2507

Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* **18**: 2529-2545

Greenland P, O'Malley PG (2005) When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. *Arch Intern Med* **165**: 2454-2456

Greenland S, Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* **142**: 1255-1264

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**: 29-36

Harel O, Zhou XH (2007) Multiple imputation: review of theory, implementation and software. *Stat Med* **26**: 3057-3077

Harrell FE (2001) Regression modelling strategies with application to linear models, logistic regression, and survival analysis. Springer-Verlag: New York

Harrell FE (2008) Design: Design Package URL:.http://stat.ethz.ch/CRAN/

Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA (1984) Regression modelling strategies for improved prognostic prediction. *Stat Med* **3**: 143-152

Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* **15**: 361-387

Harrell FE, Lee KL, Matchar DB, Reichert TA (1985) Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep* **69**: 1071-1077

Harrell FE, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, Muhe L, Gatchalian S, Eichenwald HF (1998) Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants. WHO/ARI Young Infant Multicentre Study Group. *Stat Med* **17**: 909-944

Hastie T, Sleeper L, Tibshirani R (1992) Flexible covariate effects in the proportional hazards model. *Breast Cancer Res Treat* **22**: 241-250

Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, Nicholson RI, Griffiths K (1982) A prognostic index in primary breast cancer. *Br J Cancer* **45**: 361-366

Hayden JA, Cote P, Steenstra IA, Bombardier C (2008) Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. *J Clin Epidemiol* **61**: 552-560

Heinzl H, Tempfer C (2001) A cautionary note on segmenting a cyclical covariate by minimum P-value search. *computational statistics and data analysis* **35**: 451-461

Heitjan DF, Little RJA (1991) Multiple imputation for the fatal accident reporting system. *Applied Statistics* **40**: 13-29

Hess KR (1995) Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Stat Med* **14**: 1707-1723

Heymans MW, Van Buuren S, Knol DL, Van Mechelen W, de Vet HC (2007) Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med Res Methodol* **7**: 33

Hilsenbeck SG, Clark GM (1996) Practical p-value adjustment for optimally selected cutpoints. *Stat Med* **15**: 103-112

Hilsenbeck SG, Clark GM, McGuire WL (1992) Why do so many prognostic factors fail to pan out? *Breast Cancer Res Treat* **22**: 197-206

Hollander N, Schumacher M (2001) On the problem of using 'optimal' cutpoints in the assessment of quantitative prognostic factors. *Onkologie* **24**: 194-199

Hollander N, Schumacher M (2006) Estimating the functional form of a continuous covariate's on survival time. *computational statistics and data analysis* **50**: 1131-1151

Horton T (2007) maxstat: Maximally Selected Rank Statistics URL:.http://stat.ethz.ch/CRAN/

Hosmer DW, Lemeshow S (2000) Applied logistic regression.

Hothorn T, Lausen B, Benner A, Radespiel-Troger M (2004) Bagging survival trees. *Stat Med* **23**: 77-91

Houssami N, Cuzick J, Dixon JM (2006) The prevention, detection, and management of breast cancer. *Med J Aust* **184**: 230-234

Hukkelhoven CW, Rampen AJ, Maas AI, Farace E, Habbema JD, Marmarou A, Marshall LF, Murray GD, Steyerberg EW (2006) Some prognostic models for traumatic brain injury were not valid. *J Clin Epidemiol* **59**: 132-143

Joseph L, Belisle P, Tamim H, Sampalis JS (2004) Selection bias found in interpreting analyses with missing data for the prehospital index for trauma. *J Clin Epidemiol* **57**: 147-153

Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. *Ann Intern Med* **130**: 515-524

Kirkegaard T, Bartlett JM (2006) Novel pharmacodiagnostics in breast cancer. *European oncological disease* 53-56

Kirkegaard T, McGlynn LM, Campbell FM, Muller S, Tovey SM, Dunne B, Nielsen KV, Cooke TG, Bartlett JM (2007) Amplified in breast cancer 1 in human epidermal growth factor receptor - positive tumors of tamoxifen-treated breast cancer patients. *Clin Cancer Res* **13**: 1405-1411

Kirkegaard T, Witton CJ, McGlynn LM, Tovey SM, Dunne B, Lyon A, Bartlett JM (2005) AKT activation predicts outcome in breast cancer patients treated with tamoxifen. *J Pathol* **207**: 139-146

Klein JP, Moeschberger ML (2003) Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag: New York

Kneipp SM, McIntosh M (2001) Handling missing data in nursing research with multiple imputation. *Nurs Res* **50**: 384-389

Knorr KL, Hilsenbeck SG, Wenger CR, Pounds G, Oldaker T, Vendely P, Pandian MR, Harrington D, Clark GM (1992) Making the most of your prognostic factors: presenting a more accurate survival model for breast cancer patients. *Breast Cancer Res Treat* **22**: 251-262

Kollias J, Murphy CA, Elston CW, Ellis IO, Robertson JF, Blamey RW (1999) The prognosis of small primary breast cancers. *Eur J Cancer* **35**: 908-912

Kristman VL, Manno M, Cote P (2005) Methods to account for attrition in longitudinal data: do they work? A simulation study. *Eur J Epidemiol* **20**: 657-662

Laupacis A, Sekar N, Stiell IG (1997) Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* **277**: 488-494

Lausen B, Horton T, Bertz F, Schumacher M (2004) Assessment of optimal selected prognostic factors. *Biometrical Journal* **46**: 364-374

Lausen B, Schumacher M (1992) Maximally selected rank statistics. *Biometrics* **48**: 73-85

Lee AH, Ellis IO (2008) The nottingham prognostic index for invasive carcinoma of the breast. *Pathol Oncol Res* **14**: 113-115

Linderholm B, Grankvist K, Wilking N, Johansson M, Tavelin B, Henriksson R (2000) Correlation of vascular endothelial growth factor content with recurrences, survival, and first relapse site in primary node-positive breast carcinoma after adjuvant treatment. *J Clin Oncol* **18**: 1423-1431

Lumley.T. (2008) mitools: Tools for multiple imputation of missing data URL:.http://stat.ethz.ch/CRAN/

Lundin J, Lehtimaki T, Lundin M, Holli K, Elomaa L, Turpeenniemi-Hujanen T, Kataja V, Isola J, Joensuu H (2006) Generalisability of survival estimates for patients with breast cancer--a comparison across two population-based series. *Eur J Cancer* **42**: 3228-3235

MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. *Psychol Methods* **7**: 19-40

Marshall G, Grover FL, Henderson WG, Hammermeister KE (1994) Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery. *Stat Med* **13**: 1501-1511

Marshall G, Henderson WG, Moritz TE, Shroyer AL, Grover FL, Hammermeister KE (1995) Statistical methods and strategies for working with large data bases. *Med Care* **33**: OS35-OS42

Martin M (2006) Molecular biology of breast cancer. *Clin Transl Oncol* **8**: 7-14

Mazumdar M, Glassman JR (2000) Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med* **19**: 113-132

Mazumdar M, Smith A, Bacik J (2003) Methods for categorizing a prognostic variable in a multivariable setting. *Stat Med* **22**: 559-571

McGlynn LM, Kirkegaard T, Edwards J, Tovey S, Cameron D, Twelves C, Bartlett JM, Cooke TG (2009) Ras/Raf-1/MAPK pathway mediates response to tamoxifen but not chemotherapy in breast cancer patients. *Clin Cancer Res* **15**: 1487-1495

McPherson K, Steel CM, Dixon JM (2000) ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. *BMJ* **321**: 624-628

Micheli A, Mugno E, Krogh V, Quinn MJ, Coleman M, Hakulinen T, Gatta G, Berrino F, Capocaccia R (2002) Cancer prevalence in European registry areas. *Ann Oncol* **13**: 840-865

Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. (2006) Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* **59**: 1092-1101

Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, Kane JP, Pankow JS, Devlin JJ, Willerson JT, Boerwinkle E (2007) Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *Am J Epidemiol* **166**: 28-35

Musil CM, Warner CB, Yobas PK, Jones SL (2002) A comparison of imputation techniques for handling missing data. *West J Nurs Res* **24**: 815-829

Nardi A, Schemper M (2003) Comparing Cox and parametric models in clinical studies. *Stat Med* **22**: 3597-3610

National Heart Long and Blood institute (2009) Framingham Heart Study. URL: http://www.framinghamheartstudy.org/about/index.html

Novins DK, Beals J, Mitchell CM (2001) Sequences of substance use among American Indian adolescents. *J Am Acad Child Adolesc Psychiatry* **40**: 1168-1174

O'Reilly SM, Camplejohn RS, Barnes DM, Millis RR, Rubens RD, Richards MA (1990) Node-negative breast cancer: prognostic subgroups defined by tumor size and flow cytometry. *J Clin Oncol* **8**: 2040-2046

Office for National Statistics (2009) Cancer. URL: http://www.statistics.gov.uk/cci/nugget.asp?id=915

Okugawa H, Yamamoto D, Uemura Y, Sakaida N, Yamada M, Tanaka K, Kamiyama Y (2005) Prognostic factors in breast cancer: the value of the Nottingham Prognostic Index for patients treated in a single institution. *Surg Today* **35**: 907-911

Orbe J, Ferreira E, Nunez-Anton V (2002) Comparing proportional hazards and accelerated failure time models for survival analysis. *Stat Med* **21**: 3493-3510

Ottenbacher KJ, Ottenbacher HR, Tooth L, Ostir GV (2004) A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. *J Clin Epidemiol* **57**: 1147-1152

Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* **48**: 1503-1510

Pencina MJ, D'Agostino RB (2004) Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* **23**: 2109-2123

Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* **27**: 157-172

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* **159**: 882-890

R Development Core Team (2007) R: A language and environment for statistical computing URL:.http://www.R-project.org

Radespiel-Troger M, Rabenstein T, Schneider HT, Lausen B (2003) Comparison of tree-based methods for prognostic stratification of survival data. *Artif Intell Med* **28**: 323-341

Ravdin P (2005) Assessing Adjuvant Benefit: Adjuvant Decision Making in the Era of Evidence-Based Medicine and a Broad Array
of Options.

Ring BZ, Seitz RS, Beck R, Shasteen WJ, Tarr SM, Cheang MC, Yoder BJ, Budd GT, Nielsen TO, Hicks DG, Estopinal NC, Ross DT (2006) Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer. *J Clin Oncol* **24**: 3039-3047

Royston P (2004) Multiple imputation of missing values. *The Stata journal* **4**: 227-241

Royston P (2005) Multiple imputation of missing values: update. *The Stata journal* **5**: 1-14

Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious
parametric modelling (with discussion). *Applied Statistics* **43**: 429-467

Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* **25**: 127-141

Royston P, Sauerbrei W (2003) Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Stat Med* **22**: 639-659

Royston P, Sauerbrei W (2004) A new measure of prognostic separation in survival data. *Stat Med* **23**: 723-748

Royston P, Sauerbrei W (2007) Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. *computational statistics and data analysis* **51**: 4240-4253

Royston P, Sauerbrei W (2008) Multivariable Model Building A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. John Wiley: Chichester

Royston P, Sauerbrei W, Altman DG (2000) Modeling the effects of continuous risk factors. *J Clin Epidemiol* **53**: 219-221

Rubin DB (1976) Inferences and missing data. *Biometrika* **63**: 581-590

Rubin DB (1978) Multiple imputation for non response in surveys.

Sauerbrei W (1999) The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics* **48**: 313-329

Sauerbrei W, Hubner K, Schmoor C, Schumacher M (1997) Validation of existing and development of new prognostic classification schemes in node negative breast cancer. German Breast Cancer Study Group. *Breast Cancer Res Treat* **42**: 149-163

Sauerbrei W, Meier-Hirmer C, Benner A, Royston P (2006) Multivariate regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *computational statistics and data analysis* **50**: 3464-3485

Sauerbrei W, Royston P (2007) Modelling to extract more information from clinical trials data: On some roles for the bootstrap. *Stat Med* **26**: 4989-5001

Sauerbrei W, Royston P, Binder H (2007) Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med* **26**: 5512-5528

Sauerbrei W, Schumacher M (1992) A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med* **11**: 2093-2109

Schafer JL (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall: Florida

Schafer JL (1999) Multiple imputation: a primer. *Stat Methods Med Res* **8**: 3-15

Segal MR, Bloch DA (1989) A comparison of estimated proportional hazards models and regression trees. *Stat Med* **8**: 539-550

Shrive FM, Stuart H, Quan H, Ghali WA (2006) Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Med Res Methodol* **6**: 57

Sidoni A, Bellezza G, Cavaliere A, Del SR, Scheibel M, Bucciarelli E (2004) Prognostic indexes in breast cancer: comparison of the Nottingham and Adelaide indexes. *Breast* **13**: 23-27

Sposto R (2002) Cure model analysis in cancer: an application to data from the Children's Cancer Group. *Stat Med* **21**: 293-312

Stephan P (2008) Hormone Receptor Status and Diagnosis - Estogen and Progesterone. URL: http://breastcancer.about.com/od/diagnosis/p/hormone_status.htm

Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG (2003) Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol* **56**: 441-447

Steyerberg EW, Eijkemans MJ, Habbema JD (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* **52**: 935-942

Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* **54**: 774-781

Stone JC (1986) Generalized Additive Models. *Statistical sciences* **1**: 312-314

Sun GW, Shook TL, Kay GL (1996) Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* **49**: 907-916

Tableman M, Kim JS (2003) Survival Analysis Using S: Analysis of Time-to-Event Data. Chapman & Hall/CRC: New York

Tai P, Cserni G, Van de SJ, Vlastos G, Voordeckers M, Royce M, Lee SJ, Vinh-Hung V, Storme G (2005) Modeling the effect of age in T1-2 breast cancer using the SEER database. *BMC Cancer* **5**: 130

Therneau TM and Atkinson B (2009) rpart: Recursive Partitioning URL:.http://stat.ethz.ch/CRAN/

Therneau TM and Atkinson EJ (11-2-1997) An interoduction to recursive partitioning using the rpart routine.

Therneau TM, Grambsch PM (2000) Modeling Survival Data: Extending the Cox Model. Springer-Verlag .: New York

Todd JH, Dowle C, Williams MR, Elston CW, Ellis IO, Hinton CP, Blamey RW, Haybittle JL (1987) Confirmation of a prognostic index in primary breast cancer. *Br J Cancer* **56**: 489-492

Tovey SM, Dunne B, Witton CJ, Forsyth A, Cooke TG, Bartlett JM (2005) Can molecular markers predict when to implement treatment with aromatase inhibitors in invasive breast cancer? *Clin Cancer Res* **11**: 4835-4842

Tovey SM, Reeves JR, Stanton P, Ozanne BW, Bartlett JM, Cooke TG (2006) Low expression of HER2 protein in breast cancer is biologically significant. *J Pathol* **210**: 358-362

Ture M, Tokatli F, Kurt I (2009) Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 AND ID3) in determining recurrence free survival of breast cancer patients. *Expert Systems with Applications* **36**: 2017-2026

Van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* **18**: 681-694

Van Buuren S and Oudshoorn C.G.M. (2007) mice: Multivariate Imputation by Chained Equations URL:.http://stat.ethz.ch/CRAN/

Van Buuren S and Oudshoorn K (2000) Multiple imputation by chained equations: MICE V1.0 User's manual.

Van Der Heijden GJ, Donders AR, Stijnen T, Moons KG (2006) Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol* **59**: 1102-1109

Van Houwelingen HC (2000) Validation, calibration, revision and combination of prognostic survival models. *Stat Med* **19**: 3401-3415

Vickers AJ, Elkin EB, Steyerberg E (2009) Net reclassification improvement and decision theory. *Stat Med* **28**: 525-526

Vinh-Hung V, Burzykowski T, Cserni G, Voordeckers M, Van de SJ, Storme G (2003) Functional form of the effect of the numbers of axillary nodes on survival in early breast cancer. *Int J Oncol* **22**: 697-704

Vittinghoff E, McCulloch CE (2007) Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* **165**: 710-718

Ware JH (2006) The limitations of risk factors as prognostic tools. *N Engl J Med* **355**: 2615-2617

Web of Medicine (2008a) Breast Cancer: Hormone Therapy Choices. URL: http://www.webmd.com/breast-cancer/guide/hormone-therapy-choices

Web of Medicine (2008b) Breast Cancer: Hormone Therapy Overview. URL: http://www.webmd.com/breast-cancer/hormone-therapy-overview

Weber G, Vinterbo S, Ohno-Machado L (2004) Multivariate selection of genetic markers in diagnostic classification. *Artif Intell Med* **31**: 155-167

Williams BA, Mandrekar, J. N., Mandrekar, S. J., Cha, S. S., and Furth, A. F. (2006) Finding optimal cutpoints for continuous covariates with binary and time-to-event outcomes.

Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. *Circulation* **97**: 1837-1847

Wyatt JC, Altman DG (1995) Prognostic models: clinically useful or simply forgotten. *BRITISH MEDICAL JOURNAL* **311**: 1539-1541

Zhou XH, Eckert GJ, Tierney WM (2001) Multiple imputation in public health research. *Stat Med* **20**: 1541-1549

Appendix 1: Investigation of ability of risk groups derived from the UIVS and BGVS Models to predict Recurrence Free on Tamoxifen (RFoT) and Overall Survival (OS)

Ability of the risk groups derived to predict RFoT

Among 112 recorded recurrences, in 84 of these patients the patient was at that point still on tamoxifen treatment. Patients in the lowest quartile of the BGVS and UIVS indices (which are developed to predict RFS) exhibited 7-year RFoT rate of 98% (95% C.I: 96%, 100%) and 93% (95% C.I.: 83%, 100%) respectively.

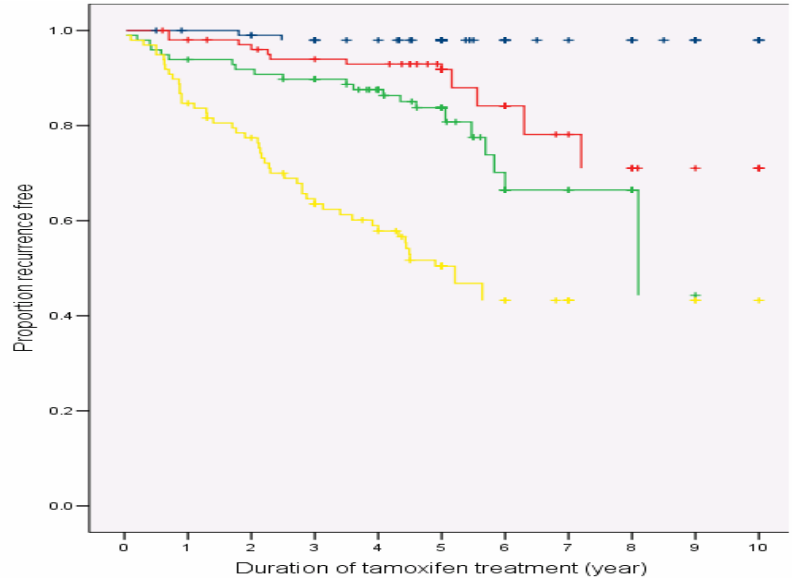
I plotted K-M curves for RFoT, using the RFS risk-groupings, to assess the extent to which the risk groups derived from BGVS and UIVS risk scores could discriminate patients with low and high risk of recurrences during tamoxifen treatment. K-M curves and estimated event-free rates indicated the ability of risk groups derived to stratify patients with respect to RFoT outcome.

Application of the BGVS and UIVS RFS risk groups to predict *RFoT*

Risk group	Index	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	BGVS	98% (96%, 100%)	**98% (96%, 100%)**	98% (96%, 100%)
	UIVS	97% (93%, 100%)	**93% (83%, 100%)**	93% (83%, 100%)
Highest	BGVS	47% (37%, 57%)	**47% (37%, 57%)**	47% (37%, 57%)
	UIVS	45% (33%, 57%)	**45% (33%, 57%)**	45% (33%, 57%)
PSEP for BGVS risk groups		51%	**51%**	51%
PSEP for UIVS risk groups		52%	**48%**	48%

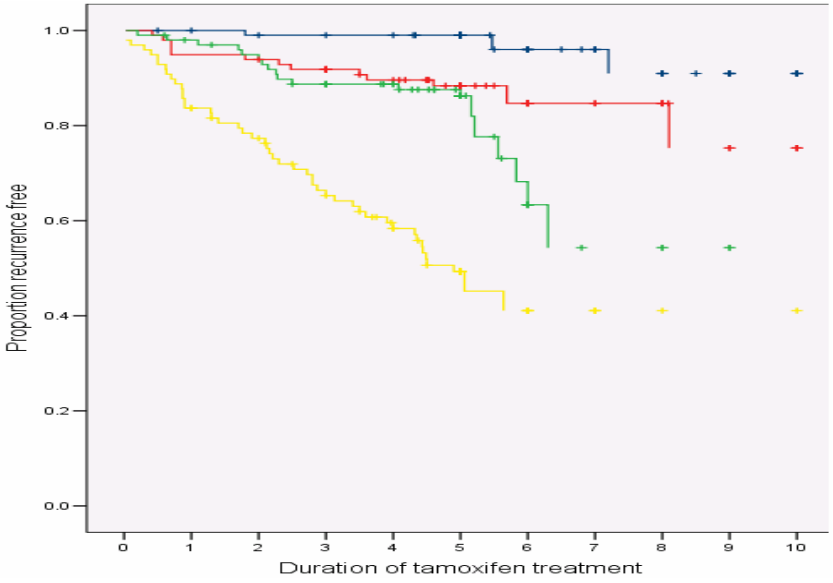
K-M curves indicating ability of the BGVS (left panel) and UIVS RFS risk groups (right panel) to predict *RFoT*

BGVS



101	95	85	18	13	8
102	92	82	12	6	5
98	86	64	10	2	1
98	60	40	9	5	3

UIVS



103	99	95	22	13	8
98	89	71	16	8	6
100	85	68	5	3	1
98	60	37	6	2	2

Ability of the risk groups derived to predict OS

By the end of follow-up there had been 74 deaths. Actuarial 7-year OS rate in the lowest risk group derived from the UIVS and BGVS indices were comparable (96% versus 98%).

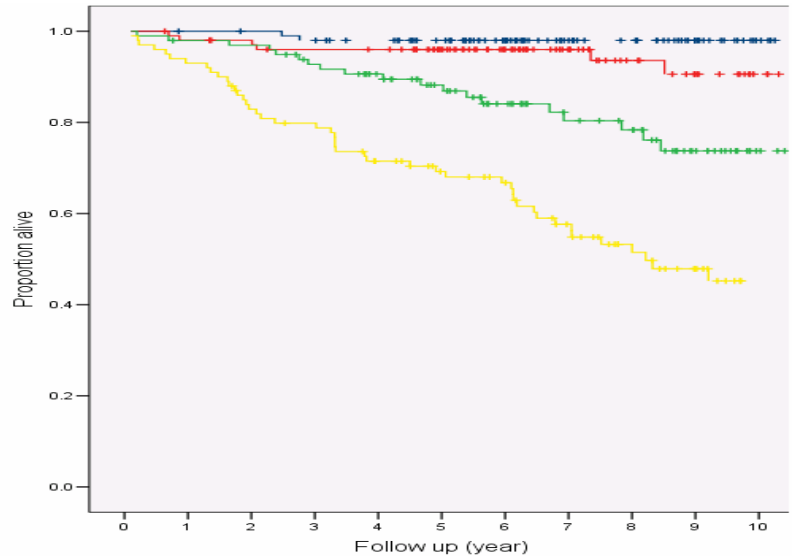
By plotting K-M curves for OS, using the RFS risk-groupings, the extent to which the risk groups derived can discriminate patients with low and high risk of death was assessed. Risk groups derived were able to stratify well diverged low and high risk patients (PSEP 47% for UIVS and 45% for BGVS).

Application of the UIVS and BGVS RFS risk groups to predict OS

Risk group	Index	5-year event free (95% C.I.)	7-year event free (95% C.I.)	10-year event free (95% C.I.)
Lowest	BGVS	98% (96%, 100%)	**98% (96%, 100%)**	98% (96%, 100%)
	UIVS	98% (96%, 100%)	**96% (92%, 100%)**	96% (92%, 100%)
Highest	BGVS	67% (57%, 77%)	**53% (43%, 63%)**	44% (32%, 56%)
	UIVS	66% (56%, 76%)	**49% (37%, 61%)**	42% (28%, 56%)
PSEP for BGVS risk groups		31%	**45%**	54%
PSEP for UIVS risk groups		32%	**47%**	54%

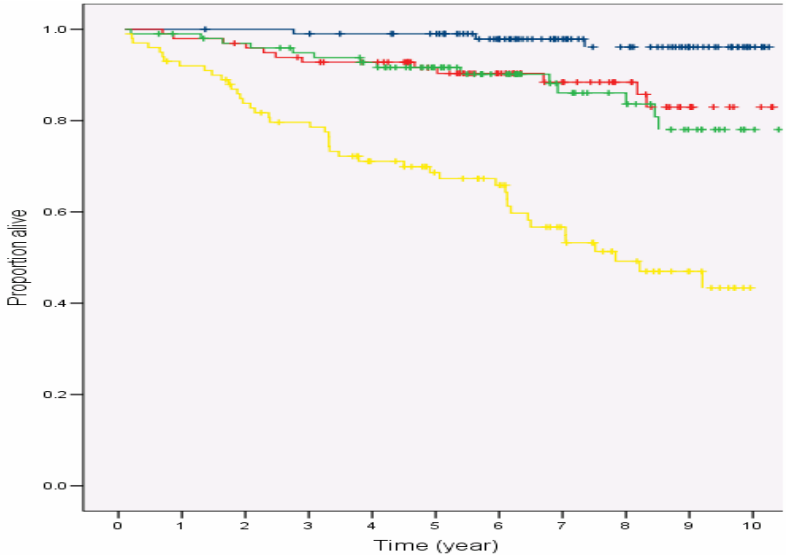
K-M curves applying the BGVS (left panel) and UIVS RFS risk groups (right panel) to predict OS

BGVS



101	97	85	50	30	18
102	94	80	48	26	17
98	87	67	43	23	10
100	77	58	41	22	12

UIVS



103	101	96	63	38	21
98	89	73	45	24	16
100	90	69	41	24	15
100	75	52	33	15	5

Appendix 2: Risk group assignment based on models developed relative to NPI

Risk groups assignment based on NPI^{q4} and the UIVS (Chapter 8)

NPI^{q4}	Status		UIVS risk groups			
			L	LI	HI	H
	Event-free	L	48	25	15	4
		LI	27	27	18	5
		HI	19	15	24	11
		H	5	9	15	22
	Recurrence	L	1	4	4	1
		LI	2	9	9	2
		HI	1	7	9	14
		H	0	2	6	41

Risk groups assignment groups based on NPI^{q4} and the BGVS (Chapter 8)

NPI^{q4}	Status		BGVS risk groups			
			L	LI	HI	H
	Event-free	L	50	27	13	2
		LI	32	23	16	6
		HI	12	26	22	9
		H	4	9	20	18
	Recurrence	L	0	4	3	3
		LI	3	7	6	6
		HI	0	3	13	15
		H	0	3	5	41

Distribution of patients into risk groups applying different imputation approaches (Chapter 9)

Median substitution	Status		MICE			
			L	LI	HI	H
	Event-free	L	15	12	0	0
		LI	15	43	20	1
		HI	0	21	40	9
		H	0	0	12	32
	Recurrence	L	2	3	0	0
		LI	2	11	6	1
		HI	0	8	17	6
		H	0	0	5	51

Distribution of patients into risk groups based on MFP and NPI^{q4} (Chapter 10)

NPI^{q4}	Status	Risk group	MFP			
			L	LI	HI	H
	Event-free	L	51	30	11	0
		LI	28	28	14	7
		HI	13	21	26	9
		H	2	6	17	26
	Recurrence	L	2	4	3	1
		LI	3	8	8	3
		HI	3	5	10	13
		H	0	1	7	41

Distribution of patients into risk groups based on linear Cox and NPI^{q4} (Chapter 10)

NPI^{q4}	Status	Risk group	Linear Cox			
			L	LI	HI	H
	Event-free	L	50	32	10	0
		LI	28	26	17	6
		HI	10	27	21	11
		H	2	3	15	31
	Recurrence	L	2	4	4	0
		LI	7	3	9	3
		HI	2	6	14	9
		H	0	0	9	40

Distribution of patients into risk groups based on Optimal split and NPI^{q4} (Chapter 10)

NPI^{q4}	Status		Optimal split			
			L	LI	HI	H
	Event-free	L	44	32	13	3
		LI	36	24	14	3
		HI	14	24	19	12
		H	2	10	16	23
	Recurrence	L	2	3	4	1
		LI	4	2	10	6
		HI	1	5	15	10
		H	0	1	6	42

Distribution of patients into risk groups based on Quartile and NPI^{q4} (Chapter 10)

NPI^{q4}	Status		Quartile			
			L	LI	HI	H
	Event-free	L	58	19	12	3
		LI	27	30	16	4
		HI	9	26	26	8
		H	2	6	17	26
	Recurrence	L	1	4	3	2
		LI	2	7	8	5
		HI	2	7	9	13
		H	0	2	8	39

Distribution of patients into risk groups based on Median and NPI^{q4} (Chapter 10)

NPI^{q4}	Status		Median			
			L	LI	HI	H
	Event-free	L	69	21	2	0
		LI	25	36	13	3
		HI	0	20	41	8
		H	0	2	11	38
	Recurrence	L	6	4	0	0
		LI	3	9	9	1
		HI	0	7	17	7
		H	0	0	6	43

Appendix 3: List of all biomarkers and univariate P-values in Fractional Polynomial (FP) analysis

Thesis abbreviation	Family	P-value	Abbreviation	Family	P-value
Krascy	RAS	<0.001	Akt1nu	AKT	0.44
Rkipnu	Non-family	<0.001	Raf1cy	MAPK	0.46
Praf338nu	MAPK	0.002	Her2me	HER	0.46
Prhisto	PgR	0.007	Hrascy	RAS	0.49
Praf338cy	MAPK	0.01	P118cy	PgR	0.49
Mapkcy	MAPK	0.01	Tace	Non-family	0.51
Pmtor	MTOR	0.02	Erbcy	PgR	0.54
Ptennu	MTOR	0.02	Akt1cy	AKT	0.54
Mtor	MTOR	0.06	H4hfr1me	HER	0.56
Akt2cy	AKT	0.06	Pher2cy	HER	0.57
Pher2nu	HER	0.07	Jrh3cy	HER	0.57
Tunel	Non-family	0.07	H4jrcy	HER	0.58
Tescy	Non-family	0.12	Pbad112c	BAD	0.60
Erbnu	PgR	0.12	Rkipcy	Non-family	0.61
Pakt1nu	AKT	0.13	H4hfr1cy	HER	0.63
Ptency	MTOR	0.14	Baxcy	BAD	0.63
P167cy	PgR	0.17	Bclxl	BAD	0.68
Ercy	PgR	0.17	Jrh3nu	HER	0.72
Erhisto	PgR	0.17	Tacep	Non-family	0.73
H4jrnu	HER	0.19	Pp70s6k3	MTOR	0.74
Praf259cy	MAPK	0.20	P118nu	PgR	0.76
P167nu	PgR	0.20	Panaktcy	AKT	0.76
Pakt2cy	AKT	0.23	Badcy	BAD	0.77
Pakt1cy	AKT	0.24	Pher2me	HER	0.81
Jrh3me	HER	0.27	Tesnu	Non-family	0.81
Panaktnu	AKT	0.27	Bcl2	BAD	0.82
Mapknu	MAPK	0.29	P167me	PgR	0.89
P118me	PgR	0.31	Nrascy	RAS	0.89
H4jrme	HER	0.31	Hrasnu	RAS	0.91
Akt3cy	AKT	0.34	Pmapknu	MAPK	0.92
Krasnu	RAS	0.34	H4hfr1nu	HER	0.94
Pakt2nu	AKT	0.35	Her2fish	HER	0.97
Raf1nu	MAPK	0.37	Pmapkcy	MAPK	0.97
Nrasnu	RAS	0.39	Aibfis1	HER (categorical)	
Praf259nu	MAPK	0.43	Aibfis2	HER (categorical)	
Aib1	HER	0.44	Egfrmax	HER (categorical)	

Krascy, Rkipnu, and Ptennu had polynomial effects
For Pmapknu, P-value corresponding to minimum P-value method was 0.003.
For Akt1nu, P-value corresponding to non-ordinal dichotomisation was 0.003.